MONTRÉAL
CANADA
29 SEPT. - 3 OCT. 2002

CODATA 2002
18th INTERNATIONAL CONFERENCE

BOOK OF ABSTRACTS

Frontiers of Scientific and Technical Data

Montréal, Canada
29 September – 3 October 2002

Recherche, Science et Technologie Québec

CODATA Secretariat, 51 Bld de Montmorency, 75006 Paris, France
Phone + 33 1 45 25 04 96, Fax + 33 1 42 88 14 66
E-mail: codata2002@dial.oleane.com
http://www.codata.org

BOOK OF ABSTRACTS

CODATA 2002:
Frontiers of Scientific and Technical Data

Montréal, Canada
29 September – 3 October 2002

Recherche, Science et Technologie Québec

CODATA Secretariat, 51 Bld de Montmorency, 75006 Paris, France
Phone + 33 1 45 25 04 96, Fax + 33 1 42 88 14 66
E-mail: codata2002@dial.oleane.com
http://www.codata.org

CODATA 2002 International Scientific Program Committee

Co-Chairmen:
Prof. Harlan Onsrud (USA) and Dr. Gordon Wood (Canada)

Topic	Member	Country
Physical Science Data	Dr. Marcelle Gaune-Escard	France
Biological Science Data	Dr. Takashi Kunisawa	Japan
Earth and Environmental Data	Prof. Liu Chuang	China
Medical and Health Data	Dr. Elliot Siegel	USA
Social Science Data	Dr. David Johnson	USA
Informatics and Technology	Mr. Glen Newton	Canada
Data Policy	Dr. Paul F. Uhlir	USA
CODATA 2015	Dr. John R. Rodgers	Canada
Engineering Data	Dr. Aleksandr Jovanovic	Germany

Table of Contents

Welcome to CODATA 2002

The 18th International CODATA Conference — "***Frontiers of Scientific and Technical Data"*** — takes place 29 September-3 October 2002 at the Hotel Delta Centre-Ville, downtown Montreal. This four-day Conference is hosted by the Canadian and US National Committees for CODATA.

CODATA 2002 continues CODATA's 36-year tradition of serving international science by holding an open and exciting conference on the latest advances in scientific and technical data.

CODATA 2002 addresses important interdisciplinary issues in scientific and technical data management and dissemination.

CODATA 2002 features six Keynote plenary lectures by some of the most renowned scientists and data experts in the world, addressing these cross-cutting themes:

- Emerging tools and techniques for data handling
- Interoperability and data integration
- Data archiving
- Legal issues in the use of scientific and technical data
- Information economics for scientific and technical data
- Ethics in the use of scientific and technical data

CODATA 2002 involves twenty-four invited speakers from different scientific disciplines expanding on the themes presented during the plenary sessions.

CODATA 2002 involves 35 parallel sessions with approximately 190 contributed oral and poster papers covering topics such as:

Physical Science Data
Social Science Data
Biological Science Data
Informatics and Technology
Earth and Environmental Data
Data Science
Medical and Health Data
Data Policy
Technical Demonstration

CODATA 2002 is designed to excite the scientific community with new ideas in data science and expose data specialists to the complexity and variety of scientific data needs among their colleagues in other disciplines, creating an environment where ideas are exchanged, synergies emerge and partnerships begin.

CODATA Officers and Executive Committee

CODATA Officers

President: Dr. John Rumble, Jr. (1998-2002)
Vice-President: Professor Akira Tsugita (1998-2002)
Secretary General: Professor Paul G. Mezey (1998-2002)
Treasurer: Dr. Jean-Jacques Royer, (2000-2004)

Executive Committee

Dr. Heinrich Behrens (2000-2002)
Ms. Lois Blaine (1998-2002)
Dr. Abdoulaye Gaye (2000-2002)
Prof Shuichi IWATA (2000-2002)
Dr Krishan Lal (2000-2002)
Prof Steve Rossouw (1998-2002)
Prof. SUN, Honglie (1996-2002)
Dr. Vladimir S Yungman (2000-2002)
Dr. Gordon H. Wood (2000-2002)

Executive Director of CODATA: **Ms. Kathleen Cass**
CODATA Secretariat
51 Bld de Montmorency, 75006 Paris, France
Phone + 33 1 45 25 04 96 / Fax + 33 1 42 88 14 66
E-mail: codata2002@dial.oleane.com
http://www.codata.org

CODATA 2002 Sponsors

National Research Council Canada
Conseil national de recherches Canada

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

Recherche, Science et Technologie Québec

NIST National Institute of Standards and Technology

UNESCO **Publi fusion**

Keynote Abstracts

1. Preserving and Archiving of Scientific and Technical Data

Trends In Archiving Digital Data
Kevin Ashley, University of London Computer Center, UK

The scientific and technical worlds have been creating and collecting information in digital form for well over 40 years, and it is arguable that they were the first to recognise the necessity of sound infrastructures to preserve that data for future reuse, examination and criticism. But it is also true that efforts were fragmented and often discipline-specific. Digital preservation is now of concern to many; it is the cultural heritage communities, business, and governments who are setting the agenda and scoping the problem. The issues are many - who pays to keep material whose value may not be realised for many years? How do we decide what to retain if we cannot keep it all ? How do we ensure we know enough about what we have preserved to enable its future use, particularly in a discipline and possibly a culture far removed from its creators? The scientific and technical communities have solutions to offer in these areas, but they can also learn from activities elsewhere. I will draw on experiences in business, scientific and cultural worlds to illustrate shared problems and possible shared solutions to these and other challenges.

---

2. Legal Issues in Using and Sharing Scientific and Technical Data

Preserving the Positive Functions of the Public Domain in Science
Pamela Samuelson, Berkeley Center for Law and Technology, University of California at Berkeley, USA

Science has greatly benefited by the absence of intellectual property rights in data and in scientific methodologies. In recent years, intellectual property has played a greater role in scientific work. While intellectual property rights may well have a positive role to play in some fields of science, so does the public domain. This talk will discuss ongoing work exploring the positive functions of the public domain. This work may help scientists and lawyers achieve a better understanding of the circumstances under which intellectual property rights will foster science and those under which preserving the public domain will be more effective in fostering science.

---

3. Interoperability and Data Integration

Integrating Bioinformatics Data into Science: From Molecules to Biodiversity
Robert J. Robbins, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Informatics - the acquisition, management, and assessment of large (huge) amounts of data - has permeated biology. GenBank contains billions of base pairs of DNA and complete genomic sequences are readily available. Microbial genomes are sequenced in a matter of days. Expression-array techniques allow the dissection of molecular function at the genomic level, while some in the biodiversity community now aspire to a global all-taxa inventory. Once, dreamers thought about assembling all of the sequence information necessary to document an entire genome. Now it is possible to imagine bringing together all of the information necessary to describe the biosphere - past and present.

But is it possible? How vast is the challenge? Are the difficulties technical, or sociological, or semantic, or ... Most importantly of all, what could we do with all of this information? Would it - in totality - be useful in any meaningful sense? Can there ever be a biological database of everything?

4. Information Economics for Scientific and Technical Data

Economics of information services for scientific and technical data in the information age: The view from a national data center in Japan
Masamitsu Negishi, NII (National Institute of Informatics), Japan

Applications of information technology continue to spread throughout the academic and business worlds. The internet was developed and utilized originally within academia where scientists and technologists enjoyed the free exchange of scientific information with their peers. As business and entertainment uses of the web grew, approaches for controlling or restricting the flow of information more responsive to the economic needs of the business community developed. Yet the needs of the scientific community for continued easy and free exchange of information remain. This talk reviews information technology, government policy, legislation and business model issues surrounding the flow of academic information in the context of economic theories for information goods. The speaker presents an overall view of the problems based on his long experience in developing and managing database and electronic library systems at the National Institute of Informatics in Japan (formally NACSIS), a national center for scientific information. The lecture concludes with a recommended scheme for cooperative, effective usable data flows among scientists and technologists across the world.

5. Emerging Tools and Techniques for Data Handling

Text Mining - the Technology To Convert Text into Knowledge?
Stan Matwin, School of Information Technology and Engineering, University of Ottawa, Canada

In this presentation we will look at Text Mining, also known as Information Extraction: the technological solution that addresses the problem of mapping technical texts into fixed-format representations, such as database records or frames. We will define the task using real-life examples. We will take a bird's eye view of the basic text mining architecture, and discuss components of the text mining systems. We will look at the existing tools and solution providers and will discuss the limits of the technology. The talk will be illustrated with author's experience in the development of a text mining tool in genomics.

6. Ethics in the Creation and Use of Scientific and Technical Data

Ethics in the Creation and Use of Scientific and Technical Data
Prof. M.G.K. Menon, Dr. Vikram Sarabhai Distinguished Professor of Department of Space and
President, LEAD, India

Science has been moving ahead at an ever increasing rapid pace. To encourage innovation and investment, there has been increasing stress on the protection of intellectual property. The international legal system relating to patents now covers a significant part of production in diverse fields and efforts exist to extend intellectual property principles to cover all types of services, traditional knowledge and scientific and technical data in the form of data bases. Questions have been raised for some time now on what the underlying principles should be that would govern intellectual property in the area of scientific and technical data. This talk addresses the interplay between legal and economic aspects on the one hand and moral and ethical aspects on the other, particularly from the viewpoint of the advancement of science itself, which is so fundamental for progress across the total spectrum of human endeavour. Issues concerning data access by the poor and by developing countries will also be addressed along with examples illustrating the direction we need to go. Ultimately, overall human good has to be the deciding factor.

Invited Cross-Cutting Themes

1. Preserving and Archiving of Scientific and Technical Data

As the volume and use of data collected worldwide continues to expand, the effective long-term preservation of these information resources likewise increases in importance. The preservation and archiving of digital scientific and technical databases in many cases poses greater, and significantly different, challenges than those in print formats. These challenges are not only technical, but involve new scientific, financial, organizational, management, legal, and policy considerations. Moreover, although many of the challenges that require sustainable solutions are the same for digital data across all disciplines, others are distinct or unique for certain disciplines or data types. Developing countries face even greater hurdles. Addressing the many different problems in the preservation and archiving of research data successfully today will bear dividends for many generations to come; the costs of failure, though incalculable, would be profound.

---

1. The Challenge of Archiving and Preserving Remotely Sensed Data
John L. Faundeen, US Geological Survey, EROS Data Center, Sioux Falls, SD, USA

Few would question the need to archive the scientific and technical (S&T) data generated by researchers. At a minimum, the data are needed for change analysis. Likewise, most people would value efforts to ensure the preservation of the archived S&T data. Future generations will use analysis techniques not even considered today. Until recently, archiving and preserving these data were usually accomplished within existing infrastructures and budgets. As the volume of archived data increases, however, organizations charged with archiving S&T data will be increasingly challenged. The US Geological Survey has had experience in this area and has developed strategies to deal with the mountain of land remote sensing data currently being managed and the tidal wave of expected new data. The Agency has dealt with archiving issues, such as selection criteria, purging, advisory panels, and data access, and has met with preservation challenges involving photographic and digital media.

---

2. The Virtual Observatory: The Future of Data and Information Management in Astrophysics
David Schade, Canadian Astronomy Data Centre, Herzberg Institute of Astrophysics, National Research Council, Canada

The concept of a "Virtual Observatory", which would put the power of numerous ground-based and space-based observatories at the fingertips of astrophysical scientists, was once a pipe dream but is now represented by funded projects in Canada, the United States, the United Kingdom, and Europe. Astronomical data has been primarily digital for 15 years and the change from analogue (e.g. photographic plates) to digital form triggered an appreciation for the scientific value of data "archiving" and the development of astronomy data centres around the world. These facilities do much more than passively "archive" their content. They have scientific and technical staff that develop the means to add value to datasets by additional processing, they integrate datasets from different wavelength regimes with one another, they distribute those data via the web, and they actively promote the use of archival data. The next step is to federate the diverse and complimentary collections residing in data centres around the world and develop seamless means for users to simultaneously access and query multi-wavelength databases and pixels and to provide the computational resources for cross-correlation and other processing. In analogy to "the greatest encyclopedia that has ever existed" that has effectively come into being because of the internet, the Virtual Observatory will be an historic leap forward in the ability of scientists, and all human beings, to understand the universe we are part of.

3. Towards a New Knowledge of Global Climate Changes: Meteorological Data Archiving and Processing Aspects
Alexander M. Sterin, All-Russian Research Institute of Hydrometeorological Information (RIHMI-WDC), Russia

This presentation will focus on a wide range of aspects related to meteorological data utilization for getting new empirical information on climate variations. The problems of meteorological data collection, their quality assurance and control, and their archiving will be discussed.

The first and the main focus will be on the problem of environmental data archiving and preservation. The collection of Russian Research Institute for Hydrometeorological Information - World Data Center (RIHMI-WDC) is currently located on 9-track magnetic tapes. The total amount of these tapes is about 60 thousand volumes. The current archiving media are obsolete, so urgent efforts on moving the collection onto modern media are beginning.

The second focus will be on the multi-level approach in constructing the informational products based on primary meteorological observational data. This approach presumes that on the lowest level (zero level) there are raw observational data. On the next level (level number one) there are the observational data that have passed the quality check procedures. Normally, in the level one the erroneous and suspicious data are flagged. The higher levels contain the derivative, data products. It appears that most customers prefer special derivative data products that are based on the primary data and that have much easier to use formats and modest volumes, rather than the primary observational data that have more complicated formats and huge volumes. The multi-level structure of the derivatives for climate studies includes the derivatives based on observational data directly (characteristics which require the calculations based on the observational data directly), derivatives of the higher level that are based on the further generalization of products - derivatives of the lower level, and so on. Examples of such a multi-level structure of data products will be given.

The third focus will be on the cycles of data processing that are required for large, data-based climate-related projects. As a result of previous experience, it is important to preserve and to reutilize the observational data collections and to provide again the main calculations. The preservation of primary observational data is very important, because it may be necessary to recalculate the products of higher levels "from the very beginning." It appears that normally these cycles may need to be repeated once (or even more than once) per decade.

The last focus will be on the software instrumentation to obtain new information and new knowledge in climate changes. The technological aspects in processing huge volumes of data in various formats will be described.

4. Strategies for Selection and Appraisal of Scientific Data for Preservation
Seamus Ross, University of Glasgow and Principal Director ERPANET, UK

With many governments and commercial organisations creating kilometres of analogue documents every year archivists have long been confronted with the challenge of handling substantial quantities of records. Recognising the impossibility of retaining this material and documenting it in ways that would enable future users to discover and use it archivists developed the concepts of appraisal. Typically archives retain only between 5% and 10% of the records created by an organisation. Indeed, in ensuring that sufficient material is retained to provide an adequate record of our cultural, scientific, and commercial heritage effective retention and disposal strategies have proven essential. As we make the transition from a paper-based world to a digital one archivists continue to recognise the power of appraisal as they attempt to manage the increasing amounts of material created digitially. The concepts that underlie appraisal are poorly understood outside the narrow confines of the archival world, but a wider appreciation of them might bring benefits to other data creating and using communities.

Appraisal is both a technical process and an intellectual activity that requires knowledge, research, and imagination on the part of the appraiser. Appraisal, characterised at its simplest level, involves establishing the value of continuing to retain and document data or records; what administrative, evidential, informational, legal, or re-usable value does a record, document, or data set. The problem is of course compounded in the digital environment by the technical aspects of the material itself. Does technology change the processes, timeframe or relevance of appraisal? Or to paraphrase from the InterPARES Appraisal TaskForce Report (January 2001) what impact does it have on ensuring that material of 'lasting value is preserved in authentic form'.

After charting the mechanisms and processes for appraisal the paper examines how the digital environment has focused attention on establishing during the appraisal process whether or not it is feasible to maintain the authenticity and integrity of digital objects over time and what impact this has on the process and point in the life of a digital objects that it must be appraised. The paper concludes by building on this work to examine the impact of the formal process of appraisal in the archiving of scientific data sets, who should be involved and responsible for the process, what appraisal criteria might be appropriate, and at what stage in the life cycle of a digital object appraisal should be cared out.

2. Legal Issues in Using and Sharing Scientific and Technical Data

The vast increases in the production of all types of digital databases and in their transfer and use for myriad purposes raises new legal concerns and amplifies others from the old print paradigm. Paramount among these is the trend toward greater legal protection of intellectual property rights in proprietary databases and the conflicts this poses with users of data, particularly those engaged in public-interest or non-profit research and education. Other legal issues that have significant effects on the production, access, and use of scientific and technical data include the protection of privacy of human subjects, export controls and national security restrictions on data exchange and dissemination, and questions of liability. As data activities continue to increase in importance, not just for research and education but in supporting economic and social development, the issues that arise at the interface of science and law in this context will become more important to study, understand and manage effectively. Scientists especially will need to communicate effectively about these issues with their national legislatures and with intergovernmental bodies such as the World Intellectual Property Organization.

1. Search for Balance: Legal Protection for Data Compilations in the U.S.
Steven Tepp, US Copyright Office, Library of Congress, USA

The United States has a long history of providing legal protection against the unauthorized use of compilations of scientific and technical data. That protection, once broad and vigorous, is now diffuse and uncertain. In light of modern Supreme Court precedent, the U.S. Congress has struggled for several years to find the appropriate balance between providing an incentive for the creation of useful compilations of data through legal protections which allow the compiler to reap commercial benefit from his work and promoting the progress of science and useful arts by allowing researchers and scientists to have unfettered access to and use of such databases. My presentation will outline the history and current state of the legal protection afforded to databases in the United States and will then discuss the different legislative models of legal protection that have been the subject of considerable debate in the U.S. Congress in recent years.

2. Legal (dis)incentives for creating, disseminating, utilizing and sharing data for scientific and technical purposes
Kenji Naemura, Keio University, Shonan-Fujisawa Campus, Japan

While Japanese policy makers differ on practical strategies for recovery and growth after a decade of struggling economy, they all agree on a view that, for restructuring the industry in a competitive environment, more vital roles should be played by advanced S&T, as well as by improved organizational and legal schemes. It is with this view that national research institutions have undergone structural reforms, and that national universities are to follow them in a near future.

Many of the enhanced legal schemes - e.g., patents to be granted to inventions in novel areas, copyrights of digital works, and other forms of IPRs - are supposed to give incentives for S&T researchers to commercialize their results. However, some schemes - e.g., private data and security protections - may become disincentives for them to disseminate, utilize and share the results.

Thus the sui generis protection of databases introduced by the EU Directive of 1996 has raised a serious concern in the scientific community. The Science Council of Japan conducted a careful study in its subcommittee on the possible merits and demerits of introducing a similar legal protection framework in this country. Its result was published as a declaration of its 136th Assembly on October 17, 2001. It emphasized "the principle of free exchange

of views and data for scientific research and education" and, expressing its opposition against a new type of legal right in addition to the copyright, stated that caution should be exercised in dealing with the international trend toward such legislation.

There are various factors that need to be considered in evaluating the advantages and disadvantages of legal protection of S&T data. They are related to the nature of research area, the data, the originating organization, the research fund, the user and his/her purpose of use, etc. Geographical, linguistic, cultural and economical conditions should also be considered when studying the consequences. After all, any incentives for advancing S&T may not be easily translated into economic figures, but other types of contributions to the humane society must be more highly valued.

---

3. Scientific and Technical Data Policy and Management in China
Sun Honglie, Chinese Academy of Sciences, Beijing, China

The 21st century is known as an information era, in which scientific and technical data, as an important information source, will have significant effects on the social and economic development of the world. Scientific and technical data contain academic, economic, social and other values. However, the basic ways of deriving the greatest value from scientific data are not just in their creation and storage, but in their dissemination and wide application. In this regard, issues of scientific and technical data policies and management have been considered as a strategic measure in the national information system and in the scientific and technical innovation programs in China. So far, scientific and technical data policy and management in China has made progress, in which:

a) A preliminary working pattern of scientific and technical data management has been shaped-the main lead being taken by government-professional sections, with scientific institutes and universities serving a subsidiary role;
b) Digitization and networking are becoming more and more universal; and
c) Professional data management organizations are being formed and expanded.

At present, the scientific and technical data policy and management in China are mainly focused on: establishing and implementing the rules for "management and sharing of national scientific and technical data"; initiating a special project for the construction of a national scientific and technical data sharing system; and developing measures for the management of this data sharing system.

---

4. A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment
J.H. Reichman, Duke University School of Law, USA and Paul F. Uhlir, The National Academies, USA

There are a number of well-documented economic, legal, and technological efforts to privatize government-generated and commercialize government-funded scientific data in the United States that were heretofore freely available from the public domain or on an "open access" basis. If these pressures continue unabated, they will likely lead to a disruption of long-established scientific research practices and to the loss of new opportunities that digital networks and related technologies make possible. These pressures could elicit one of two types of responses. One is essentially reactive, in which the public scientific community adjusts as best it can without organizing a response to the increasing encroachment of a commercial ethos upon its upstream data resources. The other would require science policy to address the challenge by formulating a strategy that would enable the scientific community to take charge of its basic data supply and to manage the resulting research commons in ways that would preserve its public good functions without impeding socially beneficial commercial opportunities. Under the latter option, the objective would be to reinforce and recreate, by voluntary means, a public space in which the traditional sharing ethos of science can be preserved and insulated from the commodifying trends. This presentation will review some approaches that the U.S. scientific community might consider in addressing this challenge, and that could have broader applicability to scientific communities outside the United States.

3. Interoperability and Data Integration

Interoperability and data integration have become essential functions not only in scientific and technical data management and information technology, but in a wide range of areas fundamental to research and the economy. With the growing number and volume of data sources, the high-speed connectivity of the Internet, the increasing number and complexity of integrable data sources and the increasing expectations of users, resolving interoperability and data integration challenges has become a research and industry focus. Incompatibilities between data formats, software systems, methodologies, standards and models of the world continue to reduce the important flow and creation of data, information, knowledge and intellectual property. Recent techniques, initiatives, and standards have rekindled activity and progress in interoperability and data integration. XML and related solution technologies have crossed traditional domain and sector boundaries, with the promise of relief from the data gridlock.

1. Interoperability in Geospatial Web Services
Jeff de La Beaujardiere, NASA Goddard Space Flight Center, USA

This talk will outline recent work on open standards for implementing interoperable geospatial web services. Beginning in 1999, a series of Testbeds--operated by the OpenGIS Consortium (OGC), sponsored in part by US federal agencies, and involving the technical participation of industry, government and academia--has developed specifications and working implementations of geographic services to be deployed over HTTP. Pilot Projects and Technology Insertion Projects have tested and deployed these standards in real-world applications.

These information-access services can provide an additional layer of interoperability above the data search capabilities provided by National Spatial Data Infrastucture (NSDI) Clearinghouse nodes. The Web Map Service (WMS; published 2000) provides graphical renderings of geodata. The Web Feature Service (WFS; 2002) provides point, line and vector feature data encoded in the XML-based Geography Markup Language (GML; 2001). The Web Coverage Service (WCS; in preparation) provides gridded or ungridded coverage data. Additional specifications for catalog, gazetteer, and fusion services are also in progress. This talk will provide an overview of these efforts and indicate current areas of application.

2. Expanding Spatial Data Infrastructure Capabilities to Optimize Use and Sharing of Geographic Data: A Developing World Perspective
Santiago Borrero, Global Spatial Data Infrastructure (GSDI), Instituto Geografico Agustin Codazzi, Colombia

The availabity of spatial data infrastructure (SDI) capabilities at all levels, backed by international standards, guidelines and policies on access to data is needed to support human sustainable development and to derive scientific, economic and social benefits from spatial information.

In this context, this paper focuses on the need for and the current situation regarding spatial data infrastructures, in particular, from the Developing World perspective. To this end, the author (i) presents GSDI and PC IDEA aims, scope and expected contribution; and (ii) then, based on these initiatives and business plans, presents observations on the possibilities for improved data availability and interoperability. More than 50 nations are in the process of developing SDI capabilities and an increasing number of geodata related initiatives at all levels. Finally, the author evaluates the need for better cooperation and coordination among spatial data initiatives and, where feasible and convenient, integration to facilitate data access, sharing and applicability.

3. Interoperability of Biological Data Resources
Hideaki Sugawara, National Institute of Genetics, Japan

Biological data resources are composed of databases and data mining tools. The International Nucleotide Sequence Database (DDBJ /EMBL /GenBank) and homology search programs are typical resources that are indispensable to life sciences and biotechnology. In addition to these fundamental resources, a number of resources are available on the Internet, e.g., those listed in the annual as we are able to observe in the yearly database issue of the journal, Nucleic Acid Research.

Biological data objects widely span: from molecule to phenotype; from viruses to mammoth; from the bottom of the sea to outer space.

Users' profile are also wide and diverse, e.g. to find anticancer drugs from any organisms in anywhere based on crosscutting heterogeneous data resources distributed in various categories and disciplines. Users often find a novel way of utilization that the developer did not imagine. Biological data resources have been often developed ad hoc without any international guidance for the standardization resulting in heterogeneous systems. Therefore, the crosscutting is a hard task for bioinformatician. It is not practical to reform large legacy systems in accordance with a standard, even if a standard is created.

Interoperability may be a solution to provide an integrated view of heterogeneous data sources distributed in many disciplines and also in distant places. We studied Common Object Request Broker Architecture (CORBA) to find that it is quite useful to make data sources interoperable in a local area network. Nevertheless, it is not straightforward to use CORBA to integrate data resources over fire-walls. CORBA is not fire-wall friendly. Recently, XML (eXtentible Markup Language) becomes widely tested and used by so-called e-Business. XML is also extensively extended to biology. However, it is not sufficient for the interoperability of biological data resources to define Document Type Definition (DTD) or XML schema. It is because multiple groups define different XML documents for the common biological object. These heterogeneous XML documents will be made interoperable by use of SOAP (Simple Object Access Protocol), WSDL (Web Service Definition Language) and UDDI (Universal Description, Discovery and Integration). The author will introduce implementation and evaluation of these technologies in WDCM (http://wdcm.nig.ac.jp), Genome Information Broker (http://gib.genes.nig.ac.jp/) and DDBJ (http://xml.nig.ac.jp).

DDBJ: DNA Data Bank of Japan
EMBL: European Molecular Biology Laboratory
GenBank: National Center for Biotechnology Information

4. The Open Archives Initiative: A low-barrier framework for interoperability
Carl Lagoze, Department of Computer Science, Cornell University, USA

The Open Archives Initiatives Protocol for Metadata Harvesting (OAI-PMH) is the result of work in the ePrints, digital library, and museum community to develop a practical and low-barrier foundation for data interoperability. The OAI-PMH provides a method for data repositories to expose metadata in various forms about their content. Harvesters may then access this metadata to build value-added services. This talk will review the history and technology behind the OAI-PMH and describe applications that build on it.

4. Information Economics for Scientific and Technical Data

With the increasing emergence of information as a valued commodity in society, conflicts and complexities are arising in the previously assumed economic models for effective advancement of science. Scientists increasingly are viewed by commercial interests as primary markets for scientific and technical data where formerly they were viewed primarily as valued producers of such data. Governments, in their attempts to reduce research budgets, and university administrators, in their attempts to increase revenues, have increasingly encouraged scientists to license and restrict access to scientific and technical databases and services. In addition, the division between public and commercial interests is blurred by the need to draw data from multiple academic and commercial sources in order to advance knowledge in many science domains. Traditional public-good and public-interest approaches to research data activities and the established scientific mores of openness serving the advancement of science need to be explored in the context of evolving alternative information economic models.

1. Legal Protection of Databases and Science in the "European Research Area":
Economic Policy and IPR Practice in the Wake of the 1996 EC Directive
Paul A. David, Stanford University and All Souls College, Oxford

At the Lisbon Meeting of the European Council in March 2000, the member states agreed that the creation of a "European Research Area"should be a high priority goal of EU and national government policies in the coming decade. Among the policy commitments taking shape are those directed toward substantially raising the level of business R&D expenditures, not only by means of subsidies and fiscal tools (e.g., tax incentive), but also through intellectual property protections aimed at "improving the environment" for business investment in R&D. The Economic Commission of the EU currently is preparing recommendations for the implementation of IP protections in future Framework Programmes and related mechanisms that fund R&D projects, including policies affecting the use of legal protections afforded to database owners under the national implementations of the EC's Directive of March 11, 1996. This paper reviews the economics issues of IPR in databases, the judicial experience and policy pressures developing in Europe in the years following the implementations of the EC's directive. It attempts to see the likely implications these will carry for scientific research in the ERA.

2. International Protection of Non-Original Databases
Helga Tabuchi, Copyright Law Division, WIPO, Geneva, Switzerland

At the request of its member States, the International Bureau of the World Intellectual Property Organization (WIPO) commissioned external consultants to prepare economic studies on the impact of the protection of non-original databases. The studies were requested to be broad, covering not only economic issues in a narrow sense, but also social, educational and access to information issues. The consultants were furthermore expected to focus in particular on the impacts in developing, least developed and transition economies.

Five of the studies were completed in early 2002 and were submitted to the Standing Committee on Copyright and Related Rights at its seventh session in May 2002. The positions of the consultants differ significantly. The studies are available on WIPO's website at <http://www.wipo.int/eng/meetings/2002/sccr/index_7.htm>.

Most recently another consultant has been commissioned to prepare an additional study that focuses on Latin American and the Caribbean region. The study will be submitted to the Committee in due course.

3. The Digital National Framework: Underpinning the Knowledge Economy
Keith Murray, Geographic Information Strategy, Ordinance Survey, UK

Decision making requires knowledge, knowledge requires reliable information and reliable information requires data from several sources to be integrated with assurance. An underlying factor in many of these decisions is geography within an integrated geographic information infrastructure.

In Great Britain, the use of geographic information is already widespread across many customer sectors (eg central government, local authorities, land & property professionals, utilities etc) and supports many hundreds of private sector applications. An independent study in 1999 showed that £100 billion of the GB GDP per annum is underpinned by Ordnance Survey information. However little of the information that is collected, managed and used today can be easily cross referenced or interchanged, often time and labour is required which does not directly contribute to the customers project goals. Ordnance Survey's direction is driven by solving customers needs such as this.

To meet this challenge Ordnance Survey has embarked on several parallel developments to ensure that customers can start to concentrate on gaining greater direct benefits from GI. This will be achieved by making major investments in the data and service delivery infrastructure the organisation provides. Key initiatives already underway aim to establish new levels of customer care, supported by establishing a new customer friendly on-line service delivery channels. The evolving information infrastructure has been designed to meet national needs but is well placed to support wider initiatives such as the emerging European Spatial Data Infrastructure (ESDI) or INSPIRE as it is now called.

Since 1999 Ordnance Survey has been independently financed through revenues from the sale of goods. It is this freedom which is allowing the organisation to further invest surplus revenues into the development of the new infrastructure. Ordnance Survey's role is not to engage in the applications market, but to concentrate on providing a high quality spatial data infrastructure. We believe that the adoption of this common georeferencing framework will support, government, business and the citizen in making the key decisions in the future, based on joined up geographic information and thereby sound knowledge.

4. Borders in Cyberspace: Conflicting Public Sector Information Policies and their Economic Impacts
Peter Weiss, Strategic Planning and Policy Office, National Weather Service, National Oceanographic and Atmospheric Administration (NOAA), USA

Many nations are embracing the concept of open and unrestricted access to public sector information -- particularly scientific, environmental, and statistical information of great public benefit. Federal information policy in the US is based on the premise that government information is a valuable national resource and that the economic benefits to society are maximized when taxpayer funded information is made available inexpensively and as widely as possible. This policy is expressed in the Paperwork Reduction Act of 1995 and in Office of Management and Budget Circular No. A-130, “Management of Federal Information Resources.” This policy actively encourages the development of a robust private sector, offering to provide publishers with the raw content from which new information services may be created, at no more than the cost of dissemination and without copyright or other restrictions. In other countries, particularly in Europe, publicly funded government agencies treat their information holdings as a commodity to be used to generate revenue in the short-term. They assert monopoly control on certain categories of information in an attempt – usually unsuccessful – to recover the costs of its collection or creation.

Such arrangements tend to preclude other entities from developing markets for the information or otherwise disseminating the information in the public interest. The US government and the world scientific and environmental research communities are particularly concerned that such practices have decreased the availability of critical data and information. And firms in emerging information dependent industries seeking to utilize public sector information find their business plans frustrated by restrictive government data policies and other anticompetitive practices.

5. Emerging Tools and Techniques for Data Handling

All facets of the scientific and technical community are moving from an information-based platform to one that is knowledge-based. This trend increases the pressure on our being able to make the transformation from data to information to knowledge as seamlessly and efficiently as possible. Simple searches of databases to find facts followed by manual efforts to make sense of those facts no longer suffice. New methods and tools to access, interpret, analyze, visualize and disseminate data, as well as the information and the knowledge ultimately derived from them, are needed. An improved understanding is needed about these major technology requirements, how they are being addressed, and how solutions are being implemented to provide reliable support to today's scientist and engineer.

1. From GeoSpatial to BioSpatial: Managing Three-dimensional Structure Data in the Sciences
Xavier R. Lopez, Oracle Corporation

Standard relational database management technology is emerging as a critical technology for managing the large volumes of 2D and 3D vector data being collected in the geographic and life sciences. For example, database technology is playing an important role in managing the terabytes of vector information used in environmental modeling, emergency management, and wireless location-based services. In addition, three-dimensional structure information is integral to a new generation of drug discovery platforms. Three dimensional structure-based drug design helps researchers generate high-quality molecules that have better pharmacological properties. This type of rational drug design is critically dependent on the comprehensive and efficient representation of both large (macro) molecules and small molecules. The macromolecules of interest are the large protein molecules of enzymes, receptors, signal transducers, hormones, and antibodies. With the recent availability of detailed structural information about many of these macromolecule targets, drug discovery is increasingly focused toward detailed structure-based analysis of the interaction of the active regions of these large molecules with candidate small-molecule drug compounds that might inhibit, enhance, or otherwise therapeutically alter the activity of the protein target. This paper will explain the means to manage the three dimensional types from the geosciences and biosciences in object-relational database technology in order to benefit from the performance, scalability, security, and reliability of commercial software and hardware platforms. This paper will highlight recent developments in database software technologies to address the 3D requirements of the life science community.

2. Benefits and Limitations of Mega-Analysis Illustrated using the WAIS
John J. McArdle, Department of Psychology, University of Virginia, USA
David Johnson, Building Engineering and Science Talent, San Diego, CA, USA

The statistical features of the techniques of meta-analysis, based on the summary statistics from many different studies, have been highly developed and are widely used (Cook et al, 1994). However, there are some key limitations to meta-analysis, especially the necessity for equivalence of measurements and inferences about individuals from groups. These problems led us to use an approach we have termed “mega-analysis” (McArdle & Horn, 1980-1999). In this approach all raw data from separate studies are used as a collective. The techniques of mega-analysis rely on a variety of methods initially developed for statistical problems of “missing data,” “selection bias,” “factorial invariance,” “test bias,” and “multilevel analyses.” In the mega-analysis of multiple sets of raw data (a) the degree to which data from different collections can be combined is raised as a multivariate statistical question, (b) unique estimation of parameters with more breadth, precision, and reliability than can be achieved by any single study, and (c) meta-analysis results emerge as a byproduct, so the assumptions may be checked and demonstrate why a simpler meta-analysis is adequate. Mega-analysis techniques are illustrated here using a collection of data from the popular ”Wechsler Adult Intelligence Scale”(WAIS), including data from thousands of people in over 100 research studies.

3. Publication, Retrieval and Exchange of Data: an Emerging Web-based Global Solution
Henry Kehiaian, ITODYS, University of Paris 7, Paris, France

In the era of enhanced electronic communication and world-wide development of information systems, electronic publishing and the Internet offer powerful tools for the dissemination of all type of scientific information. This is now made available in electronic form primary, secondary, as well as tertiary sources. However, because of the multitude of existing physico-chemical properties and variety of modes of their presentation, the computer-assisted retrieval of the numerical values, their analysis and integration in databases is as difficult as before. Accordingly, the need to have standard data formats is more important than ever. CODATA has joined forces with IUPAC and ICSTI to develop such formats.

Three years after its establishment the IUPAC-CODATA Task Group on Standard Physico-Chemical Data Formats (IUCOSPED) has made significant progress in developing the presentation of numerical property data, as well as the relevant metadata, in standardized electronic format (SELF).

The retrieval of SELFs is possible via a web-based specialized Central Data Information Source, called DataExplorer, conceived as a portal to data sources.

An Oracle database has been designed and developed for DataExplorer at FIZ Karlsruhe, Germany. URL http://www.fiz-karlsruhe.de/dataexplorer/ ID: everyone; Password: sesame. DataExplorer is now fully operational and demonstrates the concept using 4155 Chemical components, 998 Original Data Sources, 41 Property Types, and 3805 Standard Electronic Data Files (SELF). Inclusion of additional data will be actively pursued in the future.

A link has been established from DataExplorer to one of the associated Publishers, the Data Center of the Institute of Chemical Technology, Praha, Czech Republic.

Retooling SELF in SELF-ML, an XML version of the current SELF formats, is under way.

Besides an on-line demonstration of DataExplorer from FIZ-Karlsruhe and Praha, the procedure will be illustrated by computer demonstration of two publications: (1) Vapor-Liquid Equilibrium Bibliographic Database ; (2) ELDATA, the International Electronic Journal of Physico-Chemical Data.

This Project was awarded $ 100,000 under the ICSU (International Council for Science) Grants Program 2000 for new innovative projects of high profile potential

Acknowledgments
We express our sincere thanks for the financial assistance of UNESCO and ICSU and its associated organizations, IUPAC, CODATA and ICSTI, for helpful discussions to IDF, IUCr, and CAS representatives, and for the contributions of all IUCOSPED Members and Observers, to FIZ Karlsuhe administration and its highly competent programmers, and to all the associated Publishers.

4. Creating Knowledge from Computed Data for the Design of Materials
Erich Wimmer, Materials Design s.a.r.l., France and USA

The dramatic progress in computational chemistry and materials science has made it possible to carry out 'high-throughput computations' resulting in a wealth of reliable computed data including crystallographic structures, thermodynamic and thermomechanical properties, adsorption energies of molecules on surfaces, and electronic, optical and magnetic properties. An exciting perspective comes from the application of combinatorial methodologies, which allow the generation of large sets of new compounds. High-throughput computations can be employed to obtain a range of materials properties, which can be stored together with subsequent (or parallel) experimental data. Furthermore, one can include defects such as vacancies or grain boundaries in the combinatorial space, and one can apply external pressure or stress up to extreme conditions. Convenient graphical user interfaces facilitate the construction of these systems and efficient computational methods, implemented on networked parallel computers of continuously growing computational power allow the generation of an unprecedented stream of data. This lecture will discuss experience with a technology platform, MedeA (Materials Exploration and Design Analysis), which has been developed by Materials Design with the capabilities described above in mind. using heterogeneous catalysis as an example, I will illustrate how chemical concepts can be combined with high-throughput computations to transform the computed data into information and knowledge and enable the design of novel materials.

6. Ethics in the Creation and Use of Scientific and Technical Data

The gathering and compilation of scientific and technical data and the circumstances under which those data are used are at the core of many current moral debates about the role of science in society. Such debates include control over data and information about biological materials, world species diversity, genome information for plants, animals and humans, and medical records. Other discussions are focused primarily on technologies for handling scientific and technical data, ethics in cyberspace, and such topics as whether gaps in access to scientific and technical data among those in low income and high income countries will increase and what should be done about that. Even if parties desire to act responsibly in resolving conflicts, appropriate actions frequently are not clear for the individuals or groups involved, and such conflicts may not be resolvable by resort to existing law or scientific codes of conduct. In addition, access to and use of expansive scientific and technical data sets has spread well beyond academia to a much broader range of users in non-profit groups, myriad small and large businesses, and the general consumer public. Moral stances are provided from competing perspectives in all of these arenas. The ethicist questions the underlying principles and theories upon which the moralist stands and can help the scientific community evaluate which moral causes have merit and are worthy of support.

1. Ethics and Values Relating to Scientific & Technical Data: Lessons from Chaos Theory
Joan E. Sieber, NSF

Current literature reveals manifold conflicting, shifting and cross-cutting values to be reconciled if we are to pursue intelligent, data-management policies. Projects currently underway to deal with these complexities and uncertainties suggest the inevitability of a paradigm shift. Consider, e.g., questions of what data to archive, how extensively to document it, how to maintain its accessibility despite changing software and hardware, who should have access, how to allocate the costs of sharing, and so on. Traditional normative ethical theories (e.g., utilitarianism) can suggest guiding principles, and in today's global culture, recent ethical (e.g., Rawlsian) notions such as consideration of the interests of unborn generations and of persons situated very differently from oneself suddenly have immediate practical implications. However, such traditional approaches to ethical problem solving offer little guidance for dealing with problems that are highly contextual, complex, ill-defined, dynamic and fraught with uncertainty. Narrowly defined safety issues give way to notions of the ecology of life on Earth. Minor changes can have major consequences. The stakeholders are not only scientists and engineers from one's own culture, but persons, professions, businesses and governments worldwide, as they exist today and into the future. Issues of scientific freedom and openness are in conflict with issues of intellectual property, national security, and reciprocity between organizations and nations. Ethical norms, codes, principles, theories, regulations and laws vary across cultures, and often have unintended consequences that hinder ethical problem solving. Increasingly, effective ethical problem solving depends on integration with scientific and technological theory and "know how" and empirical research on the presenting ethical problem. For example, we look increasingly to psychological theories and legal concepts for clearer notions of privacy, and to social experiments, engineering solutions and methodological innovation for ways to assure confidentiality of data. We often find that one solution does not fit all related problems.

Chaos theory has taught us principles of understanding and coping with complexity and uncertainty that are applicable to ethical problem solving of data-related issues. Implications of chaos theory are explored in this presentation, both as new tools of ethical problem solving and as concepts and principles to include in the applied ethics education of students in science and engineering.

2. Understanding and improving comparative data on science and technology
Denise Lievesley, UNESCO Institute for Statistics

Statistics can serve to benefit society, but, when manipulated politically or otherwise, may be used as instruments by the powerful to maintain the status quo or even for the purposes of oppression. Statisticians working internationally face a range of ethical problems as they try to 'make a difference' to the lives of the poorest people in the world. One of the most difficult is the dilemma between open accountability and national sovereignty (in relation to what data are collected, the methods used and who is to have access to the results).

This paper will discuss the role of the UNESCO Institute for Statistics (UIS), to explain some of the constraints under which we work, and to address principles which govern our activities. The UIS is involved in

- The collection and dissemination of cross-nationally comparable data and indicators, guardianship of these databases and support of, and consultation with, users
- The analysis and interpretation of cross-national data
- Special methodological and technical projects including the development of statistical concepts
- The development and maintenance of international classifications, and standardised procedures to promote comparability of data
- Technical capacity building and other support for users and producers of data within countries
- Establishing and sharing good practice in statistics, supporting activities which improve the quality of data and preventing the re-invention of the wheel
- Advocacy for evidence-based policies

Of these activities one of the key ones is to foster the collection of comparable data across nations, the main objectives being to enable countries to gain a greater understanding of their own situation by comparing themselves with others, thus learning from one another and sharing good practice; to permit the aggregation of data across countries to provide a global picture; and to provide information for purposes of the accountability of nations and for the assessment, development and monitoring of supra-national policies.

Denise Lievesley will discuss the consultation being carried out by the UIS to ensure that the data being collected on a cross-national basis are of relevance to national policies on science and technology. The consultation process was launched with an expert meeting where changes in science policy were debated and ways in which the UIS might monitor and measure scientific and technological activities and progress across the world were identified. A background paper was produced based on the experiences and inputs of experts from different regions and organizations, which addresses key policy issues in science and technology. The UIS will use this document as a basic reference for direct consultation with UNESCO Member States and relevant institutions. A long term strategy for the collection of science and technology data will be developed as a result of these consultations.

It is vital to build on the experience of developed countries through the important statistical activities of OECD and Eurostat but nevertheless to ensure that the collection of cross-nationally harmonised data does not distort the priorities of poorer countries. We are seeking a harmony of interests in data collection and use and the views of the participants will be sought as to how this might be achieved.

3. Ethics - An Engineers' View
Horst Kremers, Comp. Sci., Berlin, Germany

The engineering profession has long experience in developing principles for appropriate relations with clients, publishing Codes of Ethics, and developing and adhering to laws controlling the conduct of professional practice. A high demand exists in society for reliable engineering in planning, design, construction and maintenance. One of the primary objectives of an engineer's actions is to provide control over a situation by providing independent advice in conformance with moral principles in addition to sound engineering principles. In a world where life to an increasing extent depends on the reliable functioning of complex information systems and where new technical techniques emerge without chance for controlled experimentation and assessment, the need to inject ethical principles into scientific and technological decisionmaking and to fully consider the consequences of professional actions is mandatory. This presentation reviews several Code of Ethics development efforts and reflects on the Codes relative to action principles in science and technology. A potential role for CODATA is presented.

4. Ethics in Scientific and Technical Communication
Hemanthi Ranasinghe, University of Sri Jayewardenepura, Sri Lanka

Research can be described as operationally successful when the research objectives are achieved and technically successful when the researcher's understanding is enhanced, more comprehensive hypothesis are developed and lessons learned from the experience. However, research is not successful scientifically until the issues, processes and findings are made known to the scientific community. Science is not an individual experience. It is shared knowledge based on a common understanding of some aspect of the physical or social world. For that reason, the social conventions of science play an important role in establishing the reliability of scientific knowledge. If these conventions are disrupted, the quality of science can suffer. Thus, the reporting of scientific research has to be right on ethical grounds too.

General category of ethics in communication covers many things. One is Error and Negligence in Science. Some researchers may feel that the pressures on them are an inducement to haste at the expense of care. For example, they may believe that they have to do substandard work to compile a long list of publications and that this practice is acceptable. Or they may be tempted to publish virtually the same research results in two different places or publish their results in "least publishable units"—papers that are just detailed enough to be published but do not give the full story of the research project described.

Sacrificing quality to such pressures can easily backfire. A lengthy list of publications cannot outweigh a reputation for shoddy research. Scientists with a reputation for publishing a work of dubious quality will generally find that all of their publications are viewed with skepticism by their colleagues. Another vital aspect of unethical behavior in scientific communication is Misconduct in Science. This entails making up data or results (fabrication), changing or misreporting data or results (falsification), and using the ideas or words of another person without giving appropriate credit (plagiarism)-all strike at the heart of the values on which science is based. These acts of scientific misconduct not only undermine progress but the entire set of values on which the scientific enterprise rests. Anyone who engages in any of these practices is putting his or her scientific career at risk. Even infractions that may seem minor at the time can end up being severely punished. Frank and open discussion of the division of credit within research groups—as early in the research process as possible and preferably at the very beginning, especially for research leading to a published paper—can prevent later difficulties.

While misallocation of credit or errors arising from negligence-are matters that generally remain internal to the scientific community. Usually they are dealt with locally through the mechanisms of peer review, administrative action, and the system of appointments and evaluations in the research environment. But misconduct in science is unlikely to remain internal to the scientific community. Its consequences are too extreme: it can harm individuals outside of science (as when falsified results become the basis of a medical treatment), it squanders public funds, and it attracts the attention of those who would seek to criticize science. As a result, federal agencies, Congress, the media, and the courts can all get involved.

All parts of the research system have a responsibility to recognize and respond to these pressures. Institutions must review their own policies, foster awareness of research ethics, and ensure that researchers are aware of the policies that are in place. And researchers should constantly be aware of the extent to which ethically based decisions will influence their success as scientists.

7. CODATA 2015

1. Scholarly Information Architecture
Paul Ginsparg, Cornell University, USA

If we were to start from scratch today to design a quality-controlled archive and distribution system for scientific and technical information, it could take a very different form from what has evolved in the past decade from pre-existing print infrastructure. Ultimately, we might expect some form of global knowledge network for research communications. Over the next decade, there are many technical and non-technical issues to address along the way, everything from identifying optimal formats and protocols for rendering, indexing, linking, querying, accessing, mining, and transmitting the information, to identifying sociological, legal, financial, and political obstacles to realization of ideal systems. What near-term advances can we expect in automated classification systems, authoring tools, and next-generation document formats to facilitate efficient datamining and long-term archival stability? How will the information be authenticated and quality controlled? What differences should be expected in the realization of these systems for different scientific research fields? What is the proper role of governments and their funding agencies in this enterprise, and what might be the role of suitably configured professional societies? These and related questions will be considered in light of recent trends.

2. The role of scientific data in a complex world
Werner Martienssen, Physikalisches Institut der Universitaet, Frankfurt am Main , Germany

Physicists try to understand and to describe the world in terms of natural laws. These laws cover two quite different approaches in physics. First, the laws show up a mathematical structure, which in general is understood in terms of first principles, of geometrical relations and of symmetry arguments. Second, the laws contain data which are characteristic for the specific properties of the phenomena and objects. Insight into the mathematical structure aims at an understanding of the world in ever more universally applicable terms. Insight into the data shows up the magnificent diversity of the world's materials and ist behavior Whereas the description of the world in terms of a unified theory one day might be reduced to only one set of equations, the amount of data necessary to describe the phenomena of the world in their full complexity seems to be open-ended.

A unified theory has not been formulated up to now; nor can we say that our knowledge about the data would be perfect in any sense, Much has to be done, still. But being asked for, where do we expect to be in data physics and chemistry in ten to fifteen years, my answer is: We -hopefully - will be able to merge the two approaches of physics. On the basis of our understanding of Materials Science and by using the methods of computational physics we will make use both of the natural laws as well as of the complete set of known data in order to modulate, to study and to generate new materials, new properties end new phenomena.

3. Life Sciences Research in 2015

David Y. Thomas, Biochemistry Department, McGill University, Montreal, Canada

Much of the spectacular progress of life sciences research in the past 30 years has come from the application of molecular biology employing a reductionist approach with single genes, often studied in simple organisms. Now from the technologies of genomics and proteomics, scientists are deluged with increasing amounts, varieties and quality of data. The challenge is how life sciences researchers will use the data output of discovery science to formulate questions and experiments for their research and turn this into knowledge. What are the important questions? We now have the capability to answer at a profound level major biological problems of how genes function, how development of organisms is controlled, and how populations interact at the cellular, organismal and population levels. What data and what tools are needed? What skills and training will be needed for the next generation of life sciences researchers? I will discuss some of the initiatives that are planned or now underway to address these problems.

Public Lectures

Biodiversité - quelles sont les espèces, où se trouvent t'elles?
Guy Baillargeon, Agriculture and Agri-Food Canada

Les connaissances sur les espèces vivantes sont documentées dans des systèmes de classification élaborés et constamment mis-à-jour par les taxonomistes. D'autre part, la connaissance de la distribution des espèces au sein de la biosphère est encore aujourd'hui principalement dérivée de l'information associée à des spécimens conservés dans les musées et les collections d'histoire naturelle. À ceci s'ajoute pour plusieurs groupes d'organismes vivants, un grand nombre d'observations individuelles colligées par des groupes d'intérêt spécialisés (tel que dans le cas des oiseaux, par les clubs d'ornithologie). Le Réseau mondial d'information sur la biodiversité (SMIB), mieux généralement connu sous son nom anglais de 'Global Biodiversity Information Facility' (GBIF), entend favoriser l'accès à toute cette information en établissant un vaste réseau distribué de bases de données scientifiques transopérables et ouvertes à tous. Encore en début d'implantation, GBIF jouera bientôt un rôle crucial en favorisant la standardisation, la digitalisation et la dissémination de l'information scientifique relative à la biodiversité partout dans le monde. Déjà, plusieurs organisations membres de GBIF ont annoncé leur intention de s'associer pour développer un inventaire de toutes les formes de vie connues (Catalogue de la Vie) et un nombre croissant d'institutions permettent l'accès direct aux données de leurs collections par voie de requêtes distribuées. La présentation fournira des exemples de ce qu'il est déjà possible de faire en matière de transopérabilité en associant un ou plusieurs systèmes de classification avec un moteur de recherche et de cartographie automatisé interreliant les données de distribution de plusieurs millions de spécimens et d'observations fournies par des dizaines d'institutions participantes à l'un des réseaux d'information distribués qui coexistent présentement sur l'Internet.

Presentation is in French; this is the English abstract:

Biodiversity - what are the species, where are they?
Guy Baillargeon, Agriculture and Agri-Food Canada

Knowledge on living species is documented through elaborate classification systems that are constantly updated by taxonomists. Knowledge on the distribution of species in the biosphere is still today mainly derived from label information associated with specimens preserved in natural history collections. In addition, for many living organisms (such as birds), large numbers of individual observations are collected by specialized interest groups. The Global Biodiversity Information Facility (GBIF) intends to facilitate access to much of these data by establishing an interoperable, distributed network of scientific databases freely available to all. Still in its early stages, GBIF is expected to play soon a crucial role in promoting the standardization, digitization and global dissemination of the world's scientific biodiversity data within an appropriate framework for property rights and due attribution. Already, organisations associated with GBIF have announced their intention of working together towards a Catalogue of Life conceived as a knowledge set of names of all known organisms and a growing number of institutions are providing direct access to the data associated with their collections via distributed queries. Examples will be presented of what is already possible in terms of interoperability when coupling one or many classifications with an automated search and map engine that interconnects millions of distributional records provided by dozens of institutions participating to one of the many distributed biodiversity information network that coexist now on the Internet.

Visualizations of our Planet's Atmosphere, Land & Oceans
Fritz Hasler, NASA Goddard Laboratory for Atmospheres, USA

See how High-Definition Television (HDTV) is revolutionizing the way we communicate science. Go back to the early weather satellite images from the 1960s and see them contrasted with the latest US and international global satellite weather movies including hurricanes & "tornadoes". See the latest visualizations of spectacular images from NASA/NOAA remote sensing missions like Terra, GOES, TRMM, SeaWiFS, Landsat 7 including new 1 - min GOES rapid scan image sequences of Nov 9th 2001 Midwest tornadic thunderstorms. New computer software tools allow us to roam & zoom through massive global images, e.g. Landsat tours of the US, and Africa, showing desert and mountain geology as well as seasonal changes in vegetation. See dust storms in Africa and smoke plumes from fires in Mexico. Fly in and through venues using 1 m IKONOS "Spy Satellite" data. See vortexes and currents in the global oceans that bring up nutrients. See the how the ocean blooms in response to these currents and El Niño/La Niña climate changes. The presentation will be made using the latest HDTV technology from a portable computer server.

Presented by Dr. Fritz Hasler of the NASA Goddard Space Flight Center. http://Etheater.gsfc.nasa.gov

Physical Science Data

Track I-C-1: Advances in Handling Physico-Chemical Data in the Internet Era (Part 1)

Chairs: William Haynes and P. Linstrom, National Institute of Standards and Technology, USA

Modern communications and computing technology is providing new capabilities for automated data management, distribution, and analysis. For these activities to be successful, data must be characterized in a manner such that all parties will be able to locate and understand each appropriate piece of information. This session will focus on characterization of physico-chemical property data by looking at two related areas: (1) the characterization of physical systems to which data are referenced and (2) the representation of data quality. Scientists have often assessed these quantities in the context of the document in which the data are presented, something automated systems cannot do. Thus, it will be important that new data handling systems find ways to express this information by using methods that can be recognized and fully understood by all users of the data.

Many challenges are presented in both of these areas:

1. **Characterization** – A heat of reaction value, for example, may be a simple scalar number but the system to which it applies is potentially quite complex. All of the species in the reaction must be identified, along with their phases, stoichiometry, the presence of any additional species or catalysts, and the temperature and pressure.

2. **Representation** – Data quality must be expressed in such a manner that all systems handling the data can deal with it appropriately. Data quality can be considered to have two major attributes: (a) the uncertainties assigned to numerical property values and (b) data integrity in the sense that the data adhere strongly to the original source and conform to well-established database rules.

---

1. The Handling of Crystallographic Data
Brian McMahon, International Union of Crystallography, England

The Crystallographic Information File (CIF) was commissioned by the International Union of Crystallography in the late 1980s to provide a common exchange format for diffraction data and the structural models derived therefrom. It specifically addressed the requirements of an information exchange mechanism that would be portable, durable, extensible and easy to manipulate, and has won widespread acceptance as a community standard. Nowadays, CIFs are created by diffractometer software, imported and exported from structure solution, refinement and visualisation programs, and used as an electronic submission format for some structural science journals.

CIF employs simple tag-value associations in a plain ASCII file, where the meanings of the tags are stored in external reference files known as data dictionaries. These dictionaries are machine-readable (in fact conforming to the same format), and provide not only a human-readable definition of the meaning of a tag, but also machine-parsable directives specifying the type and range of permitted values, contextual validity (whether an item may appear once only or multiple times) and relationships between different items. In many ways this is similar to the separation between document instances and their structural descriptions (document type definitions or DTDs) in XML, the extensible markup language that is increasingly used for document and data handling applications. However, while many existing XML DTDs describe rather general aspects of document structure, the tags defined in CIF dictionaries detail very specific pieces of information, and leave no room for ambiguity as these items are read into and written out from a variety of software applications.

Recognised tags in CIF include not only subject-specific items (e.g. the edge lengths of a crystal unit cell) but also general tags describing the creator of the file (including address and email), its revision history, related literature citations, and general textual commentary, either for formal publication or as part of a laboratory notebook record. The objective is to capture in a single file the raw experimental data, all relevant experimental conditions, and details of subsequent processing, interpretation and comment. From a complete CIF, specialist databases harvest the material they require. While such a database might be unable to store the entire content of the source file, the IUCr encourages databases to retain deposit copies of the source or to provide links from database records to the source (for example as a supplement to a published journal article).

The richness of the tag definitions also allows automated validation of the results reported in a CIF by checking their internal consistency. At present validation software is built by hand from the published descriptions of data tags, but experiments are in hand to express the relationship between numeric tags in a fully machine-readable and executable formulation. While the CIF format is unique to crystallography (and a small number of related disciplines) it has much to contribute towards the design of similar data-handling mechanisms in other formats.

---

2. Development of KDB (Korea Thermophysical Properties Databank) and Proper Use of Data and Models in Computer Aided Process Engineering Applications

Jeong Won Kang, CAPEC, Technical University of Denmark, Denmark
Rafiqul Gani, Technical University of Denmark, Denmark
Chul Soo Lee, Korea University, Korea
Ki-Pung Yoo, Sogang University, Korea

The physical property data, equilibrium data and prediction models are essential parts of process synthesis, design, optimization and operation. Although efforts to collect and organize such data and models have been performed for decades, the demand for data models and their proper and efficient use are still growing. With the financial support of MOCIE (Ministry of Commerce, Industry and Energy) of Korea, four universities have collaborated to develop a thermophysical properties databank and enhance their capacity on experimental production. The databank (KDB) contains about 4000 pure components (hydrocarbons, polymers and electrolytes) and 5000 equilibrium data sets. Most of the data were collected along with their accuracy of measurements and/or experimental uncertainties. The data can be searched by a stand-alone program or via internet. This presentation will discuss current status and features of KDB.

In process engineering applications, selecting proper data, selecting proper model, regression of the model parameter and their proper uses are the most important aspect. CAPEC(Computer Aided Process Engineering Center, Technical University of Denmark) has been developing programs to help the proper use of thermodynamic properties data and prediction models for years. A stepwise procedure to select data sets from property databases such as KDB and CAPEC-DB , generating problem specific parameters and their proper use through appropriate property models in process engineering problems has been developed in CAPEC. The presentation will also highlight the application of property model and data in specific process engineering problems.

---

3. Reliability of Uncertainty Assignments in Generating Recommended Data from a Large Set of Experimental Physicochemical Data

Qian Dong, National Institute of Standards and Technology, Boulder, CO, USA

Experimental (raw) physicochemical property data are the fundamental building blocks for generating recommended data and for developing data prediction methods. The preparation of recommended data requires a well-designed raw data repository with complete supporting information (metadata) and reliable uncertainty assessments, a series of processes involving data normalization, standardization, and statistical analysis, as well as anomaly identification and rectification. Since there are considerable duplicate measurements in a large data collection, uncertainty assessments become a key factor in selecting high quality data among related data sets. While other information in

the database can help with the selection, the uncertainty estimates provide the most important information concerning the quality of property data. This presentation will focus on the assignment and assessment of uncertainty with a large set of experimental physicochemical property data as well as the impact of uncertainty assessments on generating recommended data.

Uncertainties represent a crucial data quality attribute. They are stored in the form of a numerical value, which is interpreted as a bias for the associated property value. The addition and subtraction of this bias from the property defines a range of values. Without uncertainties, numerical property values cannot be evaluated, while inappropriate uncertainties can also be misleading. In assessing uncertainty all potential sources of errors are propagated into the uncertainty of the property. In this process, complete information on measurement techniques, sample purity, uncertainty assessment by the investigator, and investigator's experience/records, etc. is essential in establishing uncertainties by database professionals.

Reliable provision of uncertainties for property values in databases establishes the basis for determination of recommended values. However, the process of arriving at an appropriate judgment on uncertainties is rather complex. Correct assignment of uncertainty requires highly knowledgeable and skilled data professionals, and furthermore, includes a subjective component. A large-scale data collection such as TRC SOURCE makes this sophisticated task even more demanding. A recent statistical analysis on critical constants and their uncertainties assigned in TRC SOURCE reflected the difficulty in assigning reliable uncertainties and also revealed a decisive effect of uncertainties on generating recommended values. Based on this study, a computer algorithm has been developed at NIST/TRC to systematically evaluate uncertainty assessments.

---

4. Dortmund Data Bank (DDB)- Status, Accessibility and Future Plans

Jürgen Rarey and Jürgen Gmehling, University of Oldenburg, Germany

With a view to the synthesis and design of separation processes, fitting and critical examination of model parameters used for process simulation and the development of group contribution methods 1973 a computerized data bank for phase equilibrium data was started at the University of Dortmund. While at the beginning mainly VLE data for non-electrolyte mixtures ($T_b > 0$ °C) were considered, later on also VLE (including compounds with $T_b < 0$ °C), LLE, h^E, γ^∞, azeotropic, c_P^E, SLE, v^E, adsorption equilibrium, ... data for non-electrolyte and electrolyte systems as well as pure component properties were stored in a computer readable form. This data bank (Dortmund Data Bank (DDB)) now contains nearly all worldwide available phase equilibrium data, excess properties and pure component properties.

To use the full potential of this comprehensive compilation a powerful software package was developed by DDBST GmbH (www.ddbst.de) for verifying, storing, handling and processing the various pure component and mixture data. Programs for the correlation and prediction of pure component properties, phase equilibria, excess properties as well as graphical data representation were also included.

Together with the data from the Dortmund Data bank these programs allow to analyze the real mixture behavior of a system of interest and to fit reliable model parameters (g^E-models, equations of state, group contribution methods) for the synthesis and design of chemical processes on the basis of the most actual experimental data and estimation methods.

The talk will give an overview on the development, structure and contents of the DDB and will highlight certain aspects of the accessibility and use of thermophysical data in the Internet age. Future plans concerning the development of the DDB and the software package DDBSP will be discussed.

Track I-D-1: Data On Gas Hydrates

Chair: Fedor Kuznetsov, Inst. Inorg. Chem., Novosibirsk, Russia

The session will be devoted to a discussion of the status of data on gas hydrates. It is of great interest now in many countries to find reliable and economically viable ways to use the huge resources stored in nature in the form of solid gas hydrates in permafrost areas and at the bottom of the ocean. Recovery of gas from these deposits is an extremely complicated undertaking. Exploration of the deposits, development of technologies for gas recovery, conditioning and transportation, prevention of ecological hazards – all of these problems require a great variety of different data. The session will include presentations on general problems of data collection and management as well as information on data activity in this field in different countries.

1. Gas hydrates in Siberian geological structures
Albert D. Duchkov, Institute of Geophysics SB RAS, Novosibirsk, Russia

Results of prospecting of gas hydrates accumulations in continental regions of Siberia are discussed.

In Russia, the problem of existence of gas hydrates (GH) deposits is usually discussed in the context of hydrate saturation of the Cenomanian gas pool at the Messoyaha deposit in east northern part of West Siberia. However GH were never directly observed at the Messoyaha gas deposit during 40 years of investigations. One more producing horizon has been recognized in the same geological area. This horizon is related to the Cazsalin Layer of Turonian-Coniacian age, lying above the Cenomanian deposits and having more favorable PT- conditions for hydrates formation. Analysis of specific features of geologic structure, temperature regime of the section, gas composition, mineralization of formation waters, logging data, seismic prospecting materials, and sampling suggests that gas hydrates can exist in the Cazsalin Layer of the East Messoyakha deposit. One of the possible directions of further study of genesis of natural gas hydrates and estimation of the effect of gas hydrates processes on the structure of gas deposits and gas resources is study of the hydrocarbons accumulated in the Cazsalin Layer of the East Messoyakha deposit with sampling of core by a sealed thermostatically controlled corer.

The GH accumulations were found in Lake Baikal (East Siberia). Multichannel seismic studies, performed during 1989 and 1992, have revealed in Baikal the "bottom-simulating reflector" (BSR), which gives an exact evidence of existence of the lower boundary of the GH layer. It has been established that gas hydrates are distributed in South and Central parts of Lake everywhere in places where the water depth is more than 500 meters. Four types of tectonic influence were revealed: 1) modern faults shift the BSR as they do it with usual seismic boundaries; 2) older faults shift normal reflectors, the BSR has no shifts; 3) modern faults form zones, where the BSR is destroyed; 4) the processes that proceed within older faults situated closely to the base of hydrated layer leads to undulations of the BSR. The depth of lower boundary of the GH layer in Baikal ranged from 35 to 450 m. The GH presence in the Lake Baikal sediments has been confirmed by underwater borehole BDP-97 and special geological investigations. The GH accumulations were found at the surface of bottom and in sands at the depth interval 121-161 m below bottom.

2. Gas Hydrates - Where we are now?
Yuri Makogon, Petroleum Engineering Department, Texas A&M University, USA

Gas hydrates were known for more than 200 years (1778 - Priestley). However, we have been studying industrial hydrates for about 70 years. There are more than 5000 publications related to the research on gas hydrates. We have learned some properties of hydrates formed in technological systems of production and transport of gas. We know the conditions for the formation and dissociation gas hydrates, we know the methods of removing hydrate plugs from pipelines, and the prevention methods of hydrate formation.

Natural hydrates of gas have been intensively studied over the past 30-40 years. Today we know the conditions of hydrate formation in porous media in real natural conditions, we know the regions of the world where there are hydrate deposits. Over 120 gas hydrates deposits have been discovered with the reserves of over 500 trillions cubic meters. The total potential reserves of gas in hydrates is 1.5 1016 m3.

The areas of study of gas hydrates that need to be developed include:

- Properties of hydrates and hydrate-saturated media
- Conditions of formation and dissociation of hydrates in porous media
- Effective technologies for production of gas from offshore and permafrost hydrates
- Optimum conditions for storage and transportation of gas in hydrate state
- Influence of gas hydrates on the Earth environment

3. Data on kinetics and thermodynamics of gas hydrates, application to calculations of phase formation
John A. Ripmeester, SIMS, NSC, Canada

Experimental data on gas hydrates are being produced at a rapid rate, and arise from laboratory studies, field studies and industrial laboratories, each working independently. The international hydrate community has an increasing need to access reliable data on the structural and physicochemical properties on hydrates in an efficient way. The creation of an information system covering all issues relating to hydrates is essential, as this is necessary for the prediction of hydrate occurrences, both in natural and industrial environments and the control of hydrate formation and decomposition. Ultimately this will affect our ability to predict the existence of hydrate-related hazards, to judge the potential for hydrates to contribute to the global energy supply as well as their possible influence on climate change.

4. Gas Hydrates Management Program at GTI
A. Sivaraman, Gas Technology Institute, USA

Gas hydrates are an impediment to gas flow as well as a potential energy resource. When they form inside pipelines, hydrates can slow or completely block gas flow, a significant problem for producers striving to move gas from offshore wells to onshore processing facilities. Producers, gas storage, transmission companies spend millions of dollars each year on hydrate inhibitors and other actions to help prevent hydrate formation, trying to balance cost, environmental impact, efficiency and safety. Better understanding of the mechanisms that trigger hydrate formation and dissociation could lead to creation of more effective hydrate inhibitors.

The U.S. Department of Energy, Gas Research Institute (Currently GTI) and U.S Geological Survey have documented the presence of hydrates in artic Alaska, off the U.S. Atlantic and Pacific coasts, as well as in the Gulf of Mexico and the hydrate deposits contain as much as 320,000 Tcf of natural gas compared to the current

consumption of 22.5 Tcf per year in United States. Various joint industry programs are focused in drilling and producing gas from gas hydrate fields in deep waters off the coast in US and Japan.

GTI is the premier, industry-led natural gas research and development organization in the United States, dedicated to meet current and future energy and environmental challenges. At its facilities near Chicago, Illinois, GTI has assembled state-of-the art laboratories (Laser Imaging, Acoustics and Calorimetry) operated by an expert research team that is uniquely equipped to investigate the mechanism of formation and dissociation of gas hydrates; the impact of drilling fluids, the low dosage inhibitors and anti agglomerents on the hydrates. Recent results from the facility are presented.

---

5. Computer Modeling of the Properties of Gas Hydrates - The state-of-the-art
John S. Tse, Steacie, Institute for Molecular Sciences, National Research Council of Canada

Various theoretical techniques for the modelling of the physical, thermodynamics and electronic properties of gas hydrates will be reviewed. Selected examples from recent work of the author's group will be presented. Emphasis will be placed on the prediction of the dynamic properties, occupancy, formation and dissociation mechanism of gas hydrates. Perspective on using advanced simulation method for the prediction of phase equilibria will be discussed.

---

6. Natural Gas Hydrates Studies in China
Shengbo Chen and Guangmeng Guo, Institute of Geography Sciences and Natural Resources Research, China

Natural gas hydrates studies are very important. The CODATA Task Groups on Data on Natural Gas Hydrates was newly approved in October 2000. In China, gas hydrates is a potential field for studying and exploring. The area of permafrost regions accounts for 10% of the world permafrost, especially in the mid-latitude and high-altitude mountainous regions in Qinghai-Tibet Plateau. The oil-gas resources have been confirmed by exploring in the north of Tibet Plateau. It is made clear that methane emissions and carbon dioxide uptake by observation in Qinghai-Tibet Plateau. These evidences show volumes of gas hydrate may be exist. In addition, extensive sea and long shoreline make it hopeful that began to study and explore gas hydrates. In China offshore seas, mainly in South China Sea and East China Sea, obvious signs of hydrates have been distinguished in seismic reflection profile, and high temperature of seawater and high ratio of methane in fluids can be observed. All these signs and observations indicate that it is completely possible there exists a large amount of gas hydrates in China offshore seas.

In 1990, the first experimental forming of gas hydrate was finished by composing methane and water vapor in China. Subsequently, more and more university, institute and corp. involve in gas hydrates studies, including thermodynamics of hydrate formation/decomposition, seismic observation and geochemical analysis. For example, the 9 bottom simulating reflection (BSR) was found in South China Sea, and methane contents analysis by collecting sample in East China Sea have been carried out. The information of new Earth Observation System (EOS), including EOS-MODIS and EOS-MOPITT is being applied to exploring gas hydrate in Qinghai-Tibet Plateau. The land surface temperature information in the permafrost can be retrieved by the infrared data of EOS-MODIS, and the methane emissions and carbon dioxide uptake can also gained easily by EOS-MOPITT. Actually, the high temperature of sea surface by the infrared data retrieving is consistent with the distribution of high ratio of methane in fluids in East China Sea, which proved it is possible by using EOS information.

7. State of CODATA project on information system on Gas Hydrates
Fedor A. Kuznetsov, Institute of Inorganic chemistry SB RAS, Chairman of CODATA Task Group, Russia.

Previous CODATA general assembly approved establishment of task group on Gas Hydrates data. Most authoritative specialists in field of gas hydrates were invited to be members of the group. They in total represent all major field of science and technology related to gas hydrates and most of the countries, were gas hydrates studies attract significant attention. The group has developed a concept and general recommendations on the structure of the system and requirements of data.

The system thought of as a network of independent data groups, which make their own data bases in field of their expertise. What makes this network a distributed information system is set of requirements for data management accepted by all the participants. The planned system will cover information from the following areas of science and application in relation to gas hydrates:

1. Geology
2. Geophysics
3. Geochemistry
4. Chemistry and Physics of hydrate
5. Thermodynamics
6. Kinetics of gas hydrate formation, transformation and dissociation
7. Physicochemical modeling
8. Technology of development of oil and gas
9. Technology of gas hydrates deposits development
10. Ecological impact of gas hydrates exploitation.
11. Modeling of ecology
12. Economics of gas hydrates development, recovery, transportation and use.
13. Use of gas hydrates in different sectors (fuel, chemical industry…)

By now more then a hundred groups in different countries identified by now as prospective participants of creation of the system.

Present state of the system and plans for future will be reported.

Track III-C-1: Materials Databases

Chair: B.P. Yan

1. Molten Salt Database Project: Building Information and Predicting Properties
Marcelle Gaune-Escard, Ecole Polytechnique, France

The genesis of the Molten Salt Database, realized as early as 1967 with the publication of the Molten Salt Handbook by George Janz is as relevant today as it was over 30 years ago. New high-tech applications of molten salts have emerged and the need for data is crucial for the development of new processes (pyrochemical reprocessing of nuclear fuel, nuclear reactors of new generation, elaboration of new materials, new environment-friendy energetic sorces, …).

Building a world-class critically, evaluated database is a difficult and complex process, involving considerable time and money. Ultimately, the success of the project depends on positive interactions between a diverse group of people - support staff to identify and collect relevant literature, scientists to extract and evaluate the data, database experts to design and build the necessary data architecture and interfaces, database reviewers to ensure that the database is of the highest quality, and marketing staff to ensure the widest dissemination of the database. The advent of the World Wide Web (WWW) has provided another exciting component to this paradigm - a global database structure that enables direct data deposition and evaluation by the scientific community.

Also the new concepts in engineering data information system are emerging and make it possible to merge people, computers, databases and other resources in ways that were simply never possible before.

Ongoing efforts in this respect will be described with the ultimate goal of building a Virtual Molten Salt Laboratory.

These efforts are made in parallel with our current research activities on molten salts but also in interaction with those other related actions on materials and engineering. For instance, it is also intended to adapt and apply methodologies originally used for other purposes ("human genome") to the field of molten salts., as recently demonstrated for other materials by K. Rajan at RPI, using computational "informatics" tools.

2. Development of Knowledge Base System Linked to Material Database
Yoshiyuki Kaji, Japan Atomic Energy Research Institute (JAERI), Japan
Hirokazu Tsuji, Japan Atomic Energy Research Institute (JAERI), Japan
Mitsutane Fujita and Junichi Kinugawa, National Institute for Materials Science, Japan
Kenji Yoshida and Kazuki Shimura, Japan Science and Technology Corporation, Japan
Shinichi Mashiko and Shunichi Miyagawa, Japan Nuclear Cycle Development Institute, Japan
Shuichi Iwata, University of Tokyo, Japan

The distributed material database system named 'Data-Free-Way' has been developed by four organizations (the National Institute for Materials Science, the Japan Atomic Energy Research Institute, the Japan Nuclear Cycle Development Institute, and the Japan Science and Technology Corporation) under a cooperative agreement in order to share fresh and stimulating information as well as accumulated information for the development of advanced nuclear materials, for the design of structural components, etc. In the system retrieved results are expressed as a table and/or a graph.

In order to create additional values of the system, knowledge base system, in which knowledge extracted from the material database is expressed, is planned to be developed for more effective utilization of Data-Free-Way. A standard type retrieval screen is prepared for users' convenience in Data-Free-Way. If typical retrieved results through the standard type retrieval screen are available, users do not need to retrieve the data under the same condition. Moreover, if the meaning of the retrieved results and the analyzed results are stored as knowledge, the system becomes more beneficial for many users. As the first step of the knowledge base development program, knowledge notes have been made where typical retrieved results through the standard type retrieval screen and the meaning of the retrieved results are described by each organization. XML (eXtensible Markup Language) has been adopted as the description method of the retrieved results and the meaning of them. One knowledge note described with XML is stored as one knowledge which composes the knowledge base. Knowledge notes can be made at each stage of the data retrieval, the display of the retrieved results, or the graph making. A set condition at each stage can be reproduced from the knowledge note. Storing knowledge obtained as retrieved results are described with XML. And a knowledge note can be displayed using XSL (eXtensible Style Language). Since this knowledge note is described with XML, the user can easily convert the display form of the table and the graph into the data format which the user usually uses. Moreover, additional information to the retrieved numerical values such as a unit can be easily conveyed.

This paper will describe the current status of Data-Free-Way and the description method of knowledge extracted from the material database with XML.

---

3. Activity on Materials Databases in the Society of Materials Science, Japan

Tatsuo Sakai, Ritsumeikan University, Japan
Izuru Nishikawa, Osaka University, Japan
Atsushi Sugeta, Osaka University, Japan
Toshio Shuto, Mitsubishi Research Institute Inc., Japan
Masao Sakane, Ritsumeikan University, Japan
Tatsuo Inoue, Kyoto University Sakyo-ku, Japan

The data book, consisting of Vols.1, 2 and 3, was published in 1982 by the Society of Materials Science, Japan (JSMS). Volumes 1 and 2 contained numerical data of fatigue strength of metallic materials, and Vol.3 contained graphic presentations of the data. All the data were compiled as a machine-readable database and the database was opened to use in the research and engineering applications. Furthermore, after collecting additional new data, the serial data book was also published in Vols.4 and 5 in 1992 from the same society, and these data were also compiled as the database. The CGS unit system was used in Vols.1, 2 and 3, but the SI unit system was employed in Vols.4 and 5.

In order to facilitate the useful application, both data books were combined with each other as a fully revised version, and a new data book of three volumes was published by Elsevier Science B. V. and JSMS in 1996. The database was similarly revised as a new version and it was circulated as several types of medias such as Floppy Disc, DAT-Tape domestically in Japan. These databases have been widely used in the engineering applications in Japan.

In accordance with the progress of information technology, requirements to the materials database were markedly increased in the last decade. Thus, JSMS had organized some new projects to construct two other kinds of databases in the area of materials science. The first one is the database on tensile and low-cycle fatigue properties of solders. The objective materials are Sn-37Pb and Sn-3.5Ag solders, respectively. The second one is the database on the material characteristics such as stress-strain curves and temperature dependence of heat conductivity, specific heat, elongation and Young’s modulus. These databases were circulated as CD-ROM domestically in Japan.

In the present conference, the historical scope of the database construction in JSMS and their contents are introduced together with some examples of their engineering applications performed by some research groups in JSMS. Making reference to discussions in the present conference, the authors are looking for the effective method to circulate the JSMS databases in the worldwide scale.

4. Role of MITS-NIMS to Development of Materials database
Y. Xu, J. Kinugawa and K. Yagi, National Institute for Materials Science (NIMS), Japan

Material Information Technology Station (MITS) of National Institute for Materials Science (NIMS), established in October 2001, is aimed to be a worldwide information center for materials science and engineering.

Our main activities include fact-data producing and publication, literature data acquisition, and database production. We have been continuing experiments of metal creep and fatigue for 35 years, and the data are published and distributed as NIMS Data Sheets. Besides, from this year, we start literature data acquisition on materials' structure and properties. Both of the fact-data and literature data are stored and managed as databases. We are constructing more than 10 material databases, which include polymers, metals and alloys, nuclear materials, super conducting materials, etc. Online services of these databases will be available from next April.

Being aware that a simple system with only data retrieving function can not provide enough information for material research and industrial activities, in which not only data, but also data related knowledge, and decision support function are needed, we have started several new research and development projects aiming to construct intelligent material information systems with data integration, data analysis and decision support functions.

One of our projects is to develop a material risk information platform. Basing upon material property databases, material life prediction theory, and accident information databases, this platform will provide users with material risk knowledge as well as fact data, for the purpose of safe use and correct selection of materials used for high risk equipment, for example, a power plant.

Another system under construction is a decision support system for composite material design - a composite design and property prediction system. With this system, a virtual composite can be composed with optional structure and component materials. Then some basic properties such as thermal conductivity of the composite can be evaluated according to its constitution and the properties of constituents that stored in the databases.

Track III-D-1: Physical/Chemical Data Issues

Chair: Marcelle Gaune-Escard, Ecole Polytechnique, France

1. Thermodynamic Properties and Equations of State at High Energy Densities
V. E. Fortov, Institute for High Energy Densities, Russian Academy of Sciences, Moscow, Russia

During last century the range of thermodynamic parameters was greatly broadened because of rapid development of technologies. Thermodynamic properties of matter at high pressures and temperatures are very important for fundamental researches in the fields of nuclear physics, astrophysics, thermodynamics of dense plasma. A number of applications such as nuclear fusion, thermonuclear synthesis, creation of new types of weapon, comet and meteorite hazard etc. requires knowledge of experimental data in a wide region of parameters.

Traditional way of studying of thermodynamic properties of substances at high temperatures and pressures is shock-wave experiments. During last 50 years there have been published about 15000 experimental points on shock compression, adiabatic expansion and measurements of sound velocity in shock compressed matter. These data are used to determine the numerical coefficients of general functional dependencies found from theoretical considerations in semiempirical equations of state (EOS). In this work presented are different semiempirical EOS models which are used for generalization of experimental and calculated data: from simple caloric models for organic compounds and polymer materials to complex multiphase equations of state for metals. These EOS models are valid in a wide range of phase diagram and describe experimental data with good accuracy. These models are also included into the database on shock-wave experiments with public access. The database allows one to perform calculations of EOS for large amount of substances and compare the results with experimental data in graphic form via Internet by address: http://teos.ficp.ac.ru/rusbank/.

2. Internet Chemical Directory ChIN Helps Access to Variety of Chemical - Databases on Internet
Li Xiaoxia, Institute of Process Engineering (formerly Institute of Chemical Metallurgy), Chinese Academy of Sciences, China
Li Guo, Hongwei Yang, Fengguang Nie, Zhangyuan Yang & Zhihong Xu

ChIN is a comprehensive directory of Internet chemical resources on Internet and is constructed on an information base approach other than a merely collection of chemistry related links. The daily maintenance of ChIN is done with the aid of ChIN-Manager, a specific tool based on database for maintaining flexible categories, for creating resource summary pages based on different data models. ChIN has been widely recognized in China. ChIN has been summarized as a site with a huge set of evaluated resources for chemists of all disciplines by ChemDex Plus of ChemWeb.com. ChIN has been considered also as the best Internet chemistry resources index in China by ChemDex of University of Sheffield, which is a well known web directory of chemistry in the world.

As chemical databases are the basic daily tools for chemists to get information, chemical database category is one of the most important categories in ChIN. More than 300 databases have been indexed in ChIN, covering varies databases such as bibliographic databases, chemical reactions databases, chemical catalogs, databases for material safety, databases for physical properties, databases for spectroscopy, materials databases, environmental databases, chemists phone books and so on. There is also a subcategory for selected news on the progress of major commercial chemical databases. There is a summary page for each database indexed in ChIN, summary pages for related databases that may classified into different subcategories are cross linked. Up to now, among the chemical databases indexed within ChIN, more than 70 databases can be freely accessible and over 20 databases provide free searching.

The total successful requests to ChIN is over 2.5 millions since 1998 and about half are from oversea visits. About 15% requests go to indexing pages of chemical databases in ChIN.

References
1. ChIN Page, http://www.chinweb.com.cn/, the former URL is http://chin.icm.ac.cn/
2. Xiaoxia Li, Li Guo, Suhua Huang, Zonghong Liu, Zhangyuan Yang, "Database Approach in Indexing Internet Chemical Resources", World Chemistry Congress, Chemistry by Computer, OB28, Brisbane, 1 - 6 July 2001

---

3. Graph-Theoretical Concepts and Physicochemical Data
Lionello Pogliani, Dipartimento di Chimica, Università della Calabria, Italy

The molecular polarizabilities of fifty-four organic derivatives have been optimally modeled, the induced dipole moment of another set of sixty-eight organic compounds, have, instead, been less optimally modeled. The modeling was performed by the aid of particular descriptors that have been derived by the aid of graph theoretical concepts. Till recently the starting point of these modeling strategies was the hydrogen-suppressed chemical graph and pseudograph of a molecule, which for second row atoms worked quite fine. For each type of graph or pseudograph an adjacency matrix can be written. Actually the pseudograph matrix is enough to represent mathematically either a graph or a pseudograph of a hydrogen-suppressed molecule, as it encodes not only information on single connections but also on multiple connections and self-connections, which mimic multiple bonds and non-bonding electrons. From these matrices specific theoretical graph-structural invariants can be derived, among which the molecular connectivity indices and pseudoindices. For molecules with higher-row atoms, i.e., atoms with n>2, the graph representation alone was insufficient to derive a specific invariant and use was done of atomic concepts, as there was no other way to encode the contribution of the inner-core electrons of higher-row atoms. Recently, and for the first time, inner-core electrons have been successfully 'graph' encoded by the aid of complete odd-graphs, Kp, and of the corresponding adjacency Kp-pseudograph matrix. The use of complete odd-graphs to derive graph-theoretical invariants allowed an optimal modeling of the molecular polarizabilities and a not too bad modeling of the induced dipole moment of organic derivatives of better or similar quality than the modeling achieved by MM3 calculations. Other types of 'non-pure' graph-theoretical invariants achieved less satisfactory modelings.

---

4. Progress in the Development of Combustion Kinetics Databases for Liquid Fuels
Wing Tsang, National Institute of Standards and Technology, USA

This presentation is concerned with the development of a database for the simulation of gas phase combustion. In recent years simulation have become an important tool in technology. The key for effective simulations is a reliable database of information that form the essential inputs. In the area of combustion, the complexity of the process has made necessary the building of large databases. This has been hindered by the fact that gas kinetics, the disciplinary field responsible for generating the database is still a research area. Thus there has been a need for constant upgrading. Even more serious is that most combustion is carried out with liquid fuels which are complex mixtures of intermediate sized hydrocarbons. Normal alkanes are important components and they may contain ten or more carbons.

There has been considerable recent work on the oxidation of various fuels. For one or two carbon fuels there is a state of the art database GRIMECH. There are also databases that describe the formation of PAHs and soot. A complete database should contain sufficient information that will cover the oxidation and pyrolysis reactions leading to soot formation. It should also start with some important components of liquid fuel. We have now started work in this direction using heptane as our prototypical liquid fuel. The aim is to develop the kinetics sufficiently so that they can be interfaced with the existing databases mentioned earlier.

The need is for a database specifying the thermal cracking reactions of the fuel. These can be classified as bond breaking of the fuel, decomposition of the radicals formed from bond breaking and/or radical attack, decomposition of olefins, the first stable product from radical decomposition, and finally the decomposition of the olefinic radicals. This will define the nature of the competition between oxidation and cracking and the small unsaturated species that are starting point for PAH formation and soot models.

Among the four classes of reactions, processes beginning with stable compounds are in satisfactory conditions in the sense of the availability of experimental data or methods for estimation. The technically difficult problem is the quantitative specification of the decomposition modes of the radicals. This is due to the fact that larger alkyl radicals can also isomerize. Thus for many cases it is necessary to consider the decomposition modes of all the isomerization products simultaneously. Furthermore due to their low reaction thresholds energy transfer effects must be considered. This means that reaction rates are pressure as well as temperature dependent. We will describe how this problem has been solved in the C5-C7 radicals. Finally we show how the present results lay the basis for the extension of the database too much more complex fuel mixtures.

---

5. Theoretical Models for the Computation of Thermodynamic Properties of Large Organic Molecules
Vladimir S. Yungman, Glushko Thermocenter of the Russian Academy of Sciences, Russia

The past decade has seen the development of new theoretical procedures intended for the accurate prediction of thermodynamic properties, such as entropy, heat capacity and enthalpy of formation for gases. Some of the contemporary quantum chemical approaches are highly reliable but considering the present computational capabilities these approaches can be applied to relatively small molecules. For large molecules, there remain some discrepancies between experimental thermodynamic properties and those calculated using ab initio or density functional theory (DFT) methods. The present study examines the different theoretical procedures that might be quite useful for prediction of thermodynamic properties for large molecules.

The B3LYP DFT method is cost-effective level of theory to use for determination of molecular parameters, such as geometry, vibrational frequencies and potential function for internal rotation. The accuracies of thermodynamic functions in the standard state at room temperature calculated using these molecular parameters for gaseous biphenyl, polychlorinated biphenyls, and some ethers are estimated to be 1-3 J/(K·mol) for entropy and 2-5 J/(K·mol) for heat capacity. Theoretical calculations allow also estimating the potential energy for complicated hindered internal rotations.

For large molecules, accurate methods for the estimation of the enthalpy of formation values are as yet limited by the high computational cost. We used an isodesmic scheme based on a combination of theoretical and experimental data to determine the enthalpy of formation values of polychlorinated biphenyls and ethers. The best results were obtained for processes in which the reactants and products are as similar as possible. The accuracies of the calculated enthalpies of formation values at 298.15 K are estimated to be 3-6 kJ/mol.

6. Database of Geochemical Kinetics of Minerals and Rocks

Ronghua Zhang, Shumin Hu, Xuetong Zhang and Yong Wang
Open Research Laboratory of Geochemical Kinetics, Chinese Academy of Geological Sciences, Institute of Mineral Resources, China

Data of reaction rates of minerals and rocks in waters at high temperatures T and high pressures P are important in understanding the water -rock interactions in lithosphere, and in dealing with the pollution of ground water and deep buried nuclear wastes. Reaction rates have been measured experimentally in the T range 25 to 300 °C and at various pressures. A few kinetic experiments of the mineral dissolution were performed at T above 300 °C and P higher than 22 MPa. Experiments were usually carried out using flow reactors. As operating a continuous stirred tank reactor CSTR reactor, steady state dissolution rates r (mol.sec^{-1}m^{-2}) were computed from the measured solution composition using

$$r = \frac{\Delta C_i F}{\nu_i s}$$

where ΔC_i stands for the molar concentration difference between the inlet and outlet of the ith species in solution, F represents the fluid mass flow rate, v_i refers to the stoichiometric content of i in mineral, s is the total mineral surface in the reactor (m^2). As operating a flowthrough packed bed reactor PBR, mineral particles were put inside the vertical vessel. Within the PBR, a transient material balance in a column at length Z gives:

$$D_L \frac{\partial^2 C}{\partial Z^2} - U \frac{\partial C}{\partial Z} + r = \frac{\partial C}{\partial t}$$

This model characterizes mass transfer in the axial direction in terms of an effective longitudinal diffusivity D_L that is superimposed on the plug flow velocity U. The length Z and U have been known. As measured the residence time distribution function of the flow system, we can figure out the D_L. If the boundary condition and initial condition are well known, then, the dissolution rate of the mineral is derived from the following mass balance expression for the concentration of the ith solute in a reactor cell:

$$dC_i/dt = Rate\ \nu_i\ (s/V) - \Delta C_i/t$$

where C_i is the concentration of ith species, t is the average residence time, and V is the solution volume in the pressure vessel (ml).

Recently, we measured a lot of mineral dissolution rates (carbonate, fluorite, albite, zeolite, actinolite etc.) in aqueous solutions at high T and P above the water critical point, and found the fluctuation of reaction rates occurs as crossing the critical point. And also we collect reaction rate data in the literature. We performed the geochemical kinetics data base. It includes r (rate law, rate constant k, activation energy E_a, chemical affinity A etc.), the surface nature (s, surface modification…), t (contact time, accumulation time…), mineral characters (composition, structure, occurrence, etc.), reaction system, hydrodynamic and physicochemical conditions, analytical method and equipment. The rate law is:

$$R_{net} = k_+ \Pi a_i^m (1-\exp(-A/RT))$$

where R_{net} is the net rate of reaction, k_+ is the rate constant of the forward reaction, a_i is the activity of species i in the rate determining reaction raised to some power m. Others are included, e.g., incongruent dissolution, non-linear dissolution rate, non-linear dynamics in the reaction system (if happened). This data base will also provide simulation models in predicting the water/rock interaction in nature.

Track IV-A-1: Current Trends and Challenges in Development of Engineering Materials Databases

Chair: Aleksandr Jovanovic, MPA Stuttgart, Germany

This session will provide an overview of some major issues related to performance, service and use of engineering materials databases, from the viewpoint of users and developers. The aspects of interest are, e.g., use of rapid prototyping, usability (ie. user friendliness), availability (stand-alone, LAN, Internet/Intranet), safety, reliability, etc. In particular, the issue of integrated, distributed and web-oriented databases and data warehouses will be considered. Most of these aspects require different solutions so the optimum one must be found in each case.

Related issues to be addressed include:

1. Measuring performance, service and use of software databases and data warehouses;
2. Internet and intranet databases, including implications of the technology for the 'contents' (i.e. materials data) and for the users;
3. Data vs. information vs. knowledge in engineering materials databases and data warehouses - including e.g. databases of case histories, documentation, etc.;
4. Integrated data assessment for production of higher level information - e.g. automatic definition of material laws based, e.g., on stored materials data and conventional statistics;
5. Use of intelligent methods (neural networks, machine learning, case-based reasoning, fuzzy clustering, etc.);
6. Data consistency, quality/reliability, quality assurance, certification etc. in distributed systems;
7. Future trends.

Practical applications of interest would be: large materials databases, European and international engineering materials databases, intelligent databases, corrosion/fatigue/creep databases, material testing databases, databases of certified materials data, Internet databases, materials databases in technology transfer, etc.

1. Development of a Large System of Clustered Engineering Databases for Risk-Based Life Management
A. S. Jovanovic, MPA Stuttgart

The paper describes the development of complex databases system comprising currently more than 30 single databases containing data needed for the risk-based life management of components in industrial plants. The system provides basis for the development of a new European guideline in the area of risk-based life management (RBLM), inspection (RBI) and maintenance (RBIM). Full-scale application of the concepts proposed by the guideline is essentially possible only if the issue of maximum use of available data (and consequent minimization of the need to acquire further data!), and only a modern, comprehensive, but flexible database system can provide the required solution.

The database cluster is organized as a data warehouse satisfying the needs to: (a) work in highly distributed environment, both on the developers' and on the users' end; (b) work with constantly changing database structures, updated/changed at the level of the database administrator (not developer!); (c) share common tolls and tasks across all the databases (e.g. graphics, statistical evaluation, application of data mining tools, etc.); (d) assure linking and possible integration of existing databases of older generation; (e) assure transportability of the system over a wide range of operating systems.

The paper also shows how the principles of RBLM are practically applied in a European power plant, including the implementation aspects in the "non-ideal situation" (lack of data, uncertainties, need to combine experts' opinions with results of engineering analysis, etc.).

2. Open Corrosion Expertise Access Network

W.F. Bogaerts, University of Leuven - Materials Information Processing & Corrosion Engineering Labs, Belgium
H.A. Arents, Information Architects group, Flemish Regional Government, Belgium
J.H. Zheng, University of Leuven - Materials Information Processing & Corrosion Engineering Labs, Belgium
J. Hubrecht, University of Leuven - Materials Information Processing & Corrosion Engineering Labs, Belgium
R. Cottis, UMIST - Corrosion and Protection Centre, UK

The paper will describe concepts and results from the European Commission supported "OCEAN" project (Open Corrosion Expertise Network), of which the first phase is about to be finalized during 2002. The objective of this project is the design and implementation of an open, extensible system for providing access to existing corrosion information. This will be achieved through a network of interested data providers, users and developers. Where available, existing standards and technologies will be used, with the partners developing informatics and commercial protocols to allow users single-point access to distributed data collections.

One of the major difficulties of corrosion engineering is the multi-dimensional nature of the corrosion problem. A very large number of alloys are available, and these may then be exposed to an almost infinite range of environments. Thus, although many thousands of corrosion tests have been performed and numerous papers published, it remains difficult for the individual corrosion engineer to bring together the information that is relevant to a specific situation. To some extent this problem has been tackled by centralized collections of corrosion data and abstracts. However, these are limited to published information, and tend to be rather inaccessible to potential users. The latter problem relates partly to the dedicated user interfaces that are typically used with these data collections, and partly to the commercial necessities of ensuring a reasonable return for the information providers.

The OCEAN project aims to overcome these limitations through the development of open protocols for locating, paying for, and obtaining corrosion information. In this context 'information' is used very generally, and the OCEAN system is intended to cover all sources of corrosion information including large centralized data collections, individual data collections from research projects, human expertise distilled into books and expert systems, computer-assisted learning texts, multimedia resources and access to human experts. The nominated partners in the project include representatives of several categories of information providers and users, with interest groups allowing additional organizations to participate in the project. It is a specific objective of the project that OCEAN will be open; open to all information providers to offer information, and open to all data users to obtain information. At the same time the commercial value of information will be recognized through commercial protocols, and partners in the project have particular expertise in funds transfer and electronic information systems (publishing).

The detailed specification of the OCEAN system has been one of the first tasks of the project, and the approach is based on World-Wide Web technology. The core of the OCEAN system will be an intelligent database and re-director that accepts queries in a standard form and then directs them to OCEAN data sources that are registered as having information that may be relevant to the query. The data sources respond with the data requested (or a null return) to the originator of the query. For the initial phase of the project a simple query engine is used to construct correctly formatted queries from user input, and to assemble a single response from the returns from data sources. However, it will also be possible for users to issue queries directly to the OCEAN re-director, or for alternative query engines to be used. This will allow more intelligent front-ends to be developed in due course to support less expert users, or to act as software agents for experts.

3. Use of Database Technology for Saving and Rescuing of Perishing Engineering Data and Information In Eastern Europe
L. Tóth, Bay Zoltán Institute for Logistics and Production Systems

One of the driving forces in the development of engineering science are relating to the failures took place in different engineering areas. That is why the results of the failure analysis are representing a high value of worth. Due to the development of the information technology these "local worthies" could become a tool for general access. It is obvious that the results of failure analysis contains always that information which are related to that staff where the case took place, but they contains also information for general using, which support the "thumb rule" of "learning from failures". Relating to the Central and Eastern European countries many "engineering data" (including the material data and failure case studies as well) represents only the "local worth". It is caused by minimum two facts. One of them is relating to the later application of the information technology tools for saving and rescuing of perishing engineering data and information, the other is relating to the attitude of the engineering communities in these countries. Generally it can be said that the responsible specialists for the failure case studies are belonging to the middle or aged generation having the attitudes of the 1965-75 year's of these countries. It means that this generation is not familiar with the possibilities of information technology and the failure cases are regarded as "internal business" for them. Having the new generation's ability to the modern information technology tools these obstacles can be overcome. The best solution seems to be the creation of the Internet technology based ***national failure case studies warehouses***. This database contains on the one hand the *open and general information about the failure case studies* and the other hand the "*teaching aids*" related to different type of failures including the methodological procedures of examination of the failures. Having the national case studies databases they can be joined into the network. It can only be effective way to realise it if minimum two criteria are fulfilled. One of them is relating to the unified database structure, the other is to the national language. The uniform database structure and the pilot system have to be developed by using "centralised support" (EU R&D support in Europe, or the support of the insurance companies, etc.). A Hungarian initiative will be presented which contains pp. 400 failure case studies.

4. The Background and Development of MatML, a Markup Language for Materials Property Data
E. F. Begley and Charles Sturrock, National Institute of Standards and Technology, USA

MatML is an extensible markup language (XML) for the management and exchange of materials property data. Launched in October 1999 and coordinated by the National Institute of Standards and Technology, an agency of the U.S. Department of Commerce, the MatML project has drawn upon the expertise of a cross-section of the international materials community including private industry, government laboratories, universities, standards organizations, and professional societies. The background and development of MatML will be described and will include a discussion of its features and its relationship to other scientific markup languages.

Track IV-B-1: Toward Interoperable Materials Data Systems

Chair: Yoshio Monma, Kochi University of Technology, Japan

There has been growing concern for the interoperability of factual databases in the materials database community. In order to have interoperability in the heterogeneous environment of the Internet/Intranet, we need a mechanism for sharing materials information that is not dependent upon computer systems and networking. Currently, the idea of using XML-DTD for the description of materials data is welcomed internationally. Two major activities may be identified: MatML in USA and Europe and NMC's (New Material Center) XML-DTD in Japan. Using XML/Java which supposedly allow platform independence on computer systems in development and operation, some advanced materials databases have achieved success toward being truly interoperable.

This session is intended to be a natural sequel to the June 2001 MatML Workshop held at NIST (USA) and cosponsored by the VAMAS TWA 10 (Computerized Materials Data). In this Session we want to exchange ideas and experiences in building and using materials data systems intended to be interoperable in the WWW environment.

1. Requirements for Access to Technical Data – An Industrial Perspective
Timothy M. King, LSC Group, Tamworth, UK

The ultimate objective of any collaborative venture is to share understanding. Such collaboration is the fundamental basis for all social activity. The modern-day challenge is to collaborate across the globe in an environment where change is an ever-increasing factor. The digital information revolution both fuels and offers to alleviate this challenge. However, the "Tower of Babel" remains a highly relevant parable.

Integration of computer systems is a multi-level problem. While integration is increasingly available across the foundation levels of hardware, software, user access and data, semantic integration is rarely on the basis of an explicit, agreed information model. Such models control the representation of data.

XML is now a major tool in the kit of system integrators. In order to control the content of an XML file, the necessary information model is either a DTD (Document Type Definition) or, increasingly, an XML Schema. Organisations are generating large numbers of different DTDs and XML Schemas to address the needs of individual projects.

Creating information models for integration purposes causes a great deal of pain as different organisations meet to agree and define the terminology and required information capability. The XML community is new to this challenge where as the ISO sub-committee TC184/SC4 <http://www.tc184-sc4.org/> has been working for almost twenty years to create (currently) six standards, including ISO 10303 ("Product data representation and exchange" or "STEP").

The ISO/TC184/SC4 family of information standards addresses a wide range of industrial requirements. Mature parts of the standards have delivered real business benefits to various different projects. Some challenges remain in respect of such information standards: deployment in conjunction with project management requirements; facilitation of concurrent systems engineering; adoption by Small to Medium Enterprises; security; intellectual property rights; legacy systems; and integration of multiple sources. Such requirements remain the barrier between the sources of high quality scientific and technical data and the exploitation of such data within industry.

The WWW and other communities have recognised that XML as a single prevalent representation format is not sufficient and a current hot topic is ontologies. Potentially, ontologies offer a different route to integration where unified definitions across the integration levels offer the basis for automated analysis and creation of integration solutions. However, in the short term, "ontology" is a label that is in use in too many different guises and projects such as the Standard Upper Ontology <http://suo.ieee.org/> will require further development before industry is able to effectively exploit the potential power of ontologies.

---

2. The Platform System for Federation of Materials' Data by Use of XML
Toshio Shuto, Yutaka Oyatsu, Kohmei Halada, and Hiroshi Yoshizu
Mitsubishi Research Institute, Inc., Japan

For improvement of materials database as an intelligent foundation, many databases have been developed from wide ranges of materials. However, most of them are built independently for each field of research and are just as a numerical value fact data. In reality very few are realized as a full-scale utilizable database retrieval system. Regarding material database or material data as common property, easy performance of sharing or mutual use of material database is requested along with utilization of non-material specialized field. To respond to this demand, a prototype of platform system to avail mutual use across boundaries in the field of material database was developed.

---

3. XML data-description for data-sharing of material databases
Kohmei Halada, Director of Ecomaterials Center, National Institute for Materials Science, Japan
Hiroshi Yoshizu, Ecomaterilas Center, National Institute for Materials Science, Japan
Toshio Shuto, Science and Technology Research Division, Mitsubishi Research Institute, Japan
Yoshio Monma, Kochi University of Technology, Japan

The activity of VAMAS , an international collaboration of pre-standardization of advanced material based on the agreement of Versailles summit, on XML-based data-description for data-sharing of material-database is introduced. The description consists of Kernel and Modules of each field of materials properties. The Kernel is developed by NIST, USA. The description of Modules are prepared by NMC, Japan and JRC, EU.

The background of the collaboration is followings. Databases of materials data are widely distributed all over the world. However, the common procedure to retrieve and use the data from the distributed database does not exist. For individual databases, guidelines and standardizations have been prepared such as ASTM E49 especially for materials data. In today's computerized era, further development of common or standardized procedures for the data exchange system from the viewpoint of the common platform, on which data can be treated without the regard to the structure of the original database, is required. The objective of this activity is to clarify the prerequisite for the generic platform for electrical data-sharing systems of materials data. In order to promote the data-sharing system from multi-resources of materials data where each database has its own inherent structure, it is required to prepare the common basis to retrieve, refer, link and utilize the data among them with electrical exchange.

Now the project finished the Phase 0: feasibility study on the electrical data-sharing platform of distributed materials data, and goes into Phase I: Implementations. In the Implementation's phase, trial and testing the prototype of DTD* template for existing database are subjected. (*DTD is written on the assumption of XML chosen as the result of Phase 0) - creation of prototype of DTD for several existing databases

- documentation of DTD from pre-standardized database structure such as MatML
- comparison and testing with retrieval
- clarification of the requisite of the generic DTD structure for material data

By developing this data-sharing system, various properties of materials which stored in different databases can be linked on the generic platform with the standardized template, in order to use for the life-cycle design of products from comprehensive approaches such as DfE (Design for Environment), DfS (Design for Safety), etc., used in industry.

---

4. A Prototyping of Interoperable System for Data Evaluation of Creep and Fatigue Data

Tetsuya Tsujikami, Faculty of Science and Engineering, Ryukoku University, Japan
Hiroshi Fujiwara, Dept. of Environmental Systems Engineering, Kochi University of Technology, Japan
Yoshio Monma, Dept. of Environmental Systems Engineering, Kochi University of Technology, Japan
Takeshi Horikawa, Professor and Vice President, Ryukoku University, Japan

From early stage of the computerization , creep and fatigue data have been stored in computers. So far many materials databases have been built in this area. And a number of mumerical/statistical procedures for the curve fitting for creep and fatigue have been proposed. But none of them are still interoperable. Materials data systems in the era of the Internet should have the interoperability for not only the factual data but also for data evaluation modules.

Analysis of the local data in remote computers and the verification of data evaluation methods with remote data were once considered very difficult because of the lack in the interoperability. We need two aspects the interoperability here: the description of materials data and mumerical/statistical procedures to fit the equations that show the materials properties. Under the current trend a natural choice is to use the data entities by an XML-DTD and data evaluation software written as the Java applet/servlet. On the basis of an XML-DTD developed at the New Materials Center, we have developed two materials data systems for creep and fatigue data that can be accessible via the Internet. As a prototyping we implemented a few data evaluation models for creep and fatigue. But it would be easy to add models. We also compared the difference in the performance between the two types of implementations: applet and servlet, because some of the data evaluation models require nonlinear iterative computation. A demonstration will be given in the presentation.

Track IV-B-6:
Advances in Handling Physico-Chemical Data in the Internet Era (Part 2)

Chairs: William Haynes and Peter Linstrom, National Institute of Standards and Technology, USA

Modern communications and computing technology is providing new capabilities for automated data management, distribution, and analysis. For these activities to be successful, data must be characterized in a manner such that all parties will be able to locate and understand each appropriate piece of information. This session will focus on characterization of physico-chemical property data by looking at two related areas: (1) the characterization of physical systems to which data are referenced and (2) the representation of data quality. Scientists have often assessed these quantities in the context of the document in which the data are presented, something automated systems cannot do. Thus, it will be important that new data handling systems find ways to express this information by using methods that can be recognized and fully understood by all users of the data.

Many challenges are presented in both of these areas:

1. **Characterization** – A heat of reaction value, for example, may be a simple scalar number but the system to which it applies is potentially quite complex. All of the species in the reaction must be identified, along with their phases, stoichiometry, the presence of any additional species or catalysts, and the temperature and pressure.

2. **Representation** – Data quality must be expressed in such a manner that all systems handling the data can deal with it appropriately. Data quality can be considered to have two major attributes: (a) the uncertainties assigned to numerical property values and (b) data integrity in the sense that the data adhere strongly to the original source and conform to well-established database rules.

1. Materials Data on the Internet
J. H. Westbrook, Brookline Technologies, NY, USA

The availability of the Internet has provided unprecedented opportunities for both data compilers and users. With respect to materials data we will explore:

- How do we know what is available?
- How can data be accessed, interpreted, exchanged?
- What novel modes of presentation are now available?
- What organizations are active in this field and what are their programs?
  - Professional (e.g. ASM, ASTM, NIST, NIMS, VAMAS, W3C, ...)
  - Commercial (e.g. MatWeb, MDI, CES Materials Data, Pauling File, MasterMiner, MSC.Mvision, IDES, ...)
- What improvements are needed?
- Where do we go from here and how?

Examples will be illustrated of specific materials databases available on the Internet from a variety of materials data fields:

- Fundamental data (e.g. elements from the Periodic Table, phase diagrams, crystal structures, diffusion constants, ...)
- Engineering design properties
- Environmental data
- Materials Safety data

While there is no question that large and widely varied bodies of data are accessible on the Internet, significant improvements are needed promptly, or else prospective users will become so disillusioned that they abandon electronic access for data. Among the problems that need to be addressed are:

- A well-structured on-line directory to reliable data sources should be built
- Persons or organizations posting data need be encouraged to include detailed instructions for searching for and retrieving data (a title and the URL of the homepage are not usually sufficient)
- Any on-line data site must make clear the provenance of the data shown
- Any data shown should be accompanied by full metadata for both the material whose properties are shown and for the property data themselves

---

2. Physicochemical data in Landolt-Börnstein Online
R. Poerschke, Springer-Verlag, Berlin, Germany

Nearly 120 years ago the data collection Landolt Börnstein was founded in the field of Physical chemistry. The broad scope of this expert data collection in various fields ranging from Elementary Particle Physics to Technology and the strong increase in the number of primary articles forced a transition to the open New series. New volumes are planned according to the development of new fields in science and technology, whereas the former 1th to 6th edition were planned as a closed edition.

Since 1996 CD-ROMs are produced in parallel to the printed volumes. In the year 2000 Landolt Boernstein offered free access to all volumes published until 1990. This prerelease was used heavily by the 10.000 registered test users, more than two million pages were downloaded in a short period. An electronic user survey showed that more than 80% of the users wanted to have a full electronic version of LB at their working place.

End of 2001 the complete Landolt-Börnstein collection went online. A fulltext search engine allows searches for substances and properties within all 300 LB volumes, i.e. 150.000 pages and 25.000 documents. The search can be limited to a group of Landolt-Boernstein. Specific search is possible for the fields authors, document titles and tables of contents. Simultaneous search in LB and all Springer journals is possible. Users can get automatic alerting information according to their profile of interest.

Physico chemical data are collected systematically by specialists in the field and various databases were built up. LB has excellent cooperation with several database centers. First of all they provide the raw data, which are then used by authors inside or outside of the institutions to prepare selected, evaluated and recommended data for the printed version of Landolt-Börnstein. For the electronic version additional data and references can be included. All of the material is double checked by scientists and their assistants in the Landolt-Börnstein editorial office.

Examples of physicochemical data are presented:

1) Thermodynamic data of pure substances and their mixtures: cooperation with TRC/NIST in the USA and SGTE in Europe.

2) Liquid crystals database LIQCRYST, Scidex. Development of a specific graphical structure search tool for organic substances. Of course search for CAS registry numbers molecular formula, chemical names etc. is included. For a given substance the search yields a dynamical combination of a variety of physical properties, e.g. NMR, NQR and density data.

3) High quality phase equilibrium data, i.e. phase diagrams, crystallographic and other thermodynamic data in a simple to use periodic table system.

3. Expressing Measurements and Chemical Systems for Physical Property Data
Peter Linstrom, National Institute of Standards and Technology, Gaithersburg, MD, USA

Physical property data are typically associated with a measurement of a particular chemical system in a particular state. In order for such data to be effectively utilized, both the measurement and the system must be appropriately documented. In the scientific literature, this information is often presented in great detail, while in electronic databases it is often reduced to a minimal form. For example, a scientific paper may discuss the presence or absence of impurities in a reagent, while the entry in an electronic database may simply refer to the reagent as a pure compound. For many applications such an approach is reasonable, but for others it may limit the uses to which the database can be applied.

A common response to criticisms that electronic databases lack this sort of information is to note that researchers can always refer to the original literature from which the data was abstracted. While it may not be possible to match the detail of the original literature, providing richer information in this area could provide several advantages for researchers using electronic databases. If a researcher searches a database and finds three values for a property have been measured, with two measurements being quite close to each other, the researcher may conclude that the value lies near the two measurements, discarding the third. However, if the researcher is provided with information that indicates that the third measurement was made by a more reliable method, this value may be chosen instead.

A major obstacle to providing such information in electronic form is that such work requires a grammar capable of expressing such information. Since this sort of information is not always recorded, such a grammar must allow for the ability to state that such information is unknown or only known to a limited extent.

This talk will discuss some possible approaches to improving the manner in which chemical systems and measurements are expressed in electronic form. It will include examples of problems encountered in the development in the NIST Chemistry WebBook, a web site which contains physical property data compiled from several databases.

Biological Science Data

Track I-C-2: Integrated Science for Environmental Decision-making: The Challenge for Biodiversity and Ecosystems Informatics

Chairs: Gladys Cotter, U.S. Geological Survey, USA and Bonnie Carroll, Information International Associates, USA

Introductory Context: From Local to Global: We will layout the intent and overview of the session, which is to explore the issues of turning data into a viable resource for decision-making through the development of biodiversity information infrastructures and systems. Particular emphasis will be placed on issues of obtaining, managing, accessing and using data that cross differing spatial and temporal scales. Challenges of integrating current electronic monitoring data with legacy data such as museum specimens for historical context will be addressed.

---

1. Building the US National Biological Information Infrastructure: Synergy between Regional and National Initiatives
John (Jack) Hill, Houston Advanced Research Center, USA

Information concerning biodiversity and ecosystems is critical to a wide range of scientific, educational, and government uses. However, the majority of this information is not easily accessible. In 1993, the National Research Council (NRC) published a report entitled "A Biological Survey for the Nation." The report recommended that the U.S. Department of the Interior oversee the development of a National Biotic Resource Information System. The resulting system should: 1) be a distributed federation of databases designed to make existing information more accessible, 2) develop new ways to collect and distributed data and information, as well as lead in promoting data standards, 3) support continuing state efforts to develop regional and statewide environmental databases, particularly with museums, universities and similar organizations, and 4) participate in interagency initiatives to coordinate the collection and management of biodiversity data by the federal government.

In 1994, the U.S. President signed Executive Order 12906, "Coordinating Geographic Data Acquisition and Access: the National Spatial Data Infrastructure (NSDI)." The NSDI deals with the acquisition, processing, storage, and distribution of geospatial data, and is implemented by the Federal Geographic Data Committee (FGDC). At the same time, the national biotic resource information system became the NBII (web page - http://www.nbii.gov). The NBII is implemented through the auspices of the U.S. Geological Survey (USGS). The NBII works with the FGDC to increase access and dissemination of biological geospatial data through the NBII and the NSDI. The NBII biological metadata standard, is an approved "profile" or extension of the FGDC's geospatial metadata standard.

In 1998, the Biodiversity and Ecosystems Panel of the President' s Committee of Advisors on Science And Technology (PCAST) released the report titled "Teaming With Life: Investing in Science to Understand and Use America's Living Capital". The PCAST report recommended that the federal government develop the "next generation NBII" or NBII-2. This would be accomplished through a system of nodes (interconnected entry points to the NBII). In 2001, the U.S. Congress allocated the funds for the development and promotion of the node based NBII-2.

Development and implementation of the NBII nodes is underway and is being conducted in collaboration with every sector of society. There are three types of nodes. "Regional" nodes have a geographic area of responsibility and represent a regional approach to local data, environmental issues, and data collectors. Twelve (12) regional nodes are required to cover the entire U.S. "Thematic" nodes focus on a particular biological issue (i.e., bird conservation, fisheries and aquatic resources, invasive species, urban biodiversity, wildlife disease/human health, etc.). Such

issues cross regional, national, and even international boundaries. "Infrastructure" nodes are focused on issues such as the creation, adoption, and implementation of standards through the development of common tool suites, hardware and software protocols, and geospatial technologies to achieve interoperability and transparent retrieval across the entire NBII network.

This presentation will highlight NBII development, implementation, lessons learned, and successful user applications of two regional nodes, the Southern Appalachian Information Node (SAIN) and the Central Southwest/Gulf Coast Node (CSGCN). Specific NBII applications will include multiple country-, regional-, county-, and local- (site specific) level biological, environmental, and natural resource management issues.

---

2. Building a Biodiversity Information Network in India — Biodiversity Informatics and Developing World: Status and Potentials
Vishwas Chavan and S. Rajan, National Chemical Laboratory, India

The most of the striking feature of Earth is the existence of life, and the most striking feature of life is its diversity. Biodiversity, and the ecosystems that support it, contribute trillions of dollars to national and global economies. The basis of all efforts to effectively conserve biodiversity and natural ecosystems lies in efficient access to knowledgebase on biodiversity and ecosystems resources and processes. Most of the developed countries are well ahead in the race to take advantage of new electronic information opportunities to manage and build their biodiversity knowledge bases, the recognized cornerstone for their future economic, social and environmental well being.

For developing nations, which harbors rich and diversified natural resources, much of the biodiversity information is neither available nor accessible. Hence there is a need for organized, well-resourced, national approach to build and manage biodiversity information through collaborative efforts by this group of Third World Nations.

This paper reviews the state of information technology applications in the field of biodiversity informatics in these nations, with India as model nation. India is one of the 12 mea-biodiversity countries bestowed with rich floral and faunal diversity. With its deteriorating status of natural resources and developmental activities, India is one of the best model nation for such a review. Attempts made by the author's group to develop and implement cost-efficient, easy-to-use tools for biological data management are described in brief. Feasibility of employing available tools, techniques and standards for biological data acquisition, organization, analysis, modeling and forecasting has been discussed keeping in view the informatics awareness amongst the biologists and ecologists as well as planners. With specific reference to Indian biodiversity, authors suggest the framework to build national information infrastructure to correlate, analyze and communicate biological information to help these nations to generate sustainable wealth from nature.

---

3. Developing and Integrating Data Resources from a North American Perspective
Jorge Soberon, CONABIO, Mexico

Biodiversity Information denotes a very heterogenous set of data formats, updating regimes, quality, and users. The data in the labels of biological specimens provide a natural organizing framework because the georeference and the taxonomic name can be used to link to geographically organized data (remote sensing, cartography) and to a variety of points of view (ecological or genetical data, legislation, traffic, etc.). Label data, however is widely distributed over hundreds of institutions. In this talk, we describe the technical and organizational problems that were solved to create REMIB (the World Network of Biodiversity Information), that links nearly 5 million specimens from 61 collections of 16 institutions in three countries. We also give one example of the use that such system may have.

4. Ecological Informatics: a Long-Term Ecological Research Perspective
William Michener, Long Term Ecological Research Program, USA

Scientists within the Long-Term Ecological Research (LTER) Network have provided leadership in ecological informatics since the inception of LTER in 1980. The success of LTER, where research projects span wide temporal and spatial scales, depends on the quality and longevity of the data collected. Scientists have devised data collection, data entry, data access, QA/QC and archiving strategies for ensuring that high quality data are appropriately managed to meet the needs of a broad user base for decades to come. The LTER cross-site Network Information System (NIS) is being developed to foster data sharing and collaboration among sites. Recent and important milestones for LTER include adoption of Ecological Metadata Language as a standard as well as supporting metadata software. Current and future foci include developing data standardization protocols and semantic mediation engines, both of which will facilitate LTER modeling efforts.

5. The Global Biodiversity Information Facility (GBIF) — Challenges and Opportunities from a Global Perspective
Guy Baillargeon, Agriculture and Agri-Food Canada

The Global Biodiversity Information Facility (GBIF) is a new international scientific cooperative project based on an agreement between countries, economies, and international organizations. The primary goal of GBIF is to establish an interoperable, distributed network of databases containing scientific biodiversity information in order to make the world's scientific biodiversity data freely available to all. GBIF will play a crucial role in promoting the standardization, digitization and global dissemination of the world's scientific biodiversity data within an appropriate framework for property rights and due attribution. Initially, GBIF will focus on species and specimen level data in 4 priority areas: data access and data interoperability; digitization of natural history collection data, electronic catalogue of names of known organisms; outreach and capacity building. With an expected staff of only 14, GBIF will work mostly with others in order to catalyse synergistic activities between participants, generate new investments and eliminate barriers to cooperation. In its first year of activity, GBIF has been concentrating on organisational logistics, staffing, and consultations with Scientific and Technical Advisory Groups (STAGs). Initial work plans are being drafted by the Science committee and its 4 subcommittees. Once functional, GBIF will allow to unlock and liberate vast amounts of biodiversity occurrence data for use in research and environmental decision-making. Life itself, in all its diversity (from molecules, to species, to ecosystems) will provide numerous new additional sets of data layers for integrated environmental analysis, modelling and forecasting.

Track III-C-2: Proteome Database

Chair: Akira Tsugita, Proteomics Research Laboratory, Tsukuba, Japan

Proteomics research is growing broadly and exponentially. Such research includes: extraction of protein mixture from cells and tissues, separation and isolation of the proteins (by 2-DE, HPLC etc.), and identification of the protein (by terminal sequence, in-gel digestion-MALDI-TOF-MS, Capillary-LC/ESI-MS-MS, etc). This research has goals such as: 1) Establishment of a protein catalogue, a complete list of all distinct proteins which include post-translational modification and multiple spliced variant and cleavage products. This information corresponds to genome information; 2) Correlation to protein/protein interaction; 3) Correlation to protein/nucleic acid interaction; 4) Establishment of structure/active motif information; 5) Tissue-specific protein expression; 6) Age-specific protein expression; and 7) Intra-cellular protein expression. The proteome is now applied pharmacology and medicine. Recently, the international HUPO (human proteome organisation) was established and extremely active research has been carried out. While the genome sequence is uni-dimensional and finite, the proteome information is multi-dimensional with quasi-infinite dimensions. The proteome is dynamic and constantly changing in response to various environmental factors and signals. This session is devoted to the evaluation, compilation, and dissemination of such proteome data, and to a discussion of proteome information patenting.

1. A Proteomic Approach to the Study of Cancer
Julio E Celis, Institute of Cancer Biology, Danish Cancer Society and Danish Centre for Human Genome Research, Denmark

During the past 20 years, high resolution two dimensional polyacrylamide gel electrophoresis (2D PAGE) has been the technique of choice for analysing the protein composition of cell types, tissues and fluids, as well as for studying changes in protein expression profiles elicited by various effectors. The technique, which was originally described by O'Farrell and Klose, separates proteins both in terms of their isoelectric point (pI) and molecular weight. Usually, one chooses a condition of interest and lets the cell reveal the global protein behavioral response as all detected proteins can be analyzed both qualitatively (post translational modifications) and quantitatively (relative abundance, corregulated proteins) in relation to each other [http://biobase.dk/cgi bin/celis]. Presently, high resolution 2D PAGE provides the highest resolution for protein analysis and is a key technique in proteomics, an emerging area of research of the post-genomic era that deals with the global analysis of gene expression using a plethora of technology to resolve (2D PAGE), identify (mass spectrometry, Western immunoblotting, etc.), quantitate and characterize proteins, identify interacting partners as well as to store (comprehensive 2D PAGE databases), communicate and interlink protein and DNA mapping and sequence information from ongoing genome projects. Proteomics, together with genomics, cDNA arrays, phage antibody libraries and transgenic models belong to the armamentarium of technology comprising functional genomics. Here I will report on our efforts to apply proteomic technologies to the study of bladder cancer.

2. A Proposition of XML Format for Proteomics Database
Kenichi Kamijo, T. Yamazaki and A. Tsugita, Proteomics Reseach Center, NEC Corporation, Japan

We propose XML (eXtensible Markup Language) format for proteomics database to exchange proteome analysis data. The XML-based data is highly machine-readable and easy to represent information hierarchy and relationships. There have been several XML formats of proteome data which mainly represent the siquence information stored in the Protein Indentication Resource (PIR) and the Protein Data Base(PDB).

Our XML-based data model is a proteome-analysis-oriented structure and describes information of sample preparetion, 2D gel electrophoresis images, spot identification information in the gels and the sequence information of the spots. The model is used to exchange both of preparation parameters and the results of 2D gel electrophoresis analysis. It would be accelerated collaboration among proteomics researchers if a platform exchanging these data is developed on the internet.

By using our XML-based model for proteomics, we have developed web-based prototype system which consisits of XML database, agent, security and graphical user interface(GUI).

3. Proteomics : An Important Post-genomic Tool for Understanding Gene Function
Richard J. Simpson, L. M. Connolly, D. F. Frecklington, H. Ji, G. E. Reid, M. J. Layton, and R. L. Moritz, Joint ProteomicS Laboratory (JPSL), Ludwig Institute for Cancer Research and Walter & Eliza Hall Institute for Medical Research, Melbourne, Australia

If DNA is the blueprint to build the complex machine that is a human, then proteins are the parts of the machine that make it work. With the completion of the first draft of the DNA sequence that makes up the human genome, the challenge facing medical research now is to understand gene function. Proteomics provides a biological tool, or assay, for elucidating gene function.

While the term proteomics is often synonymous with high-throughput protein profiling of normal versus diseased tissue by 2-D gel analysis, this definition is very limiting. Increasingly, the power of proteomics is being recognized for its ability to unravel intricate protein-protein interactions associated with intracellular protein trafficking and signaling pathways (i.e., cell-mapping proteomics). The technology issues associated with expression proteomics (the study of global changes in protein expression) and cell-mapping proteomics (the systematic study of protein-protein interactions through the isolation of protein complexes) are almost identical and only differ in front-end scale-up processes. The application of proteomics for studying various biological problems will be presented with representative examples of (a) differential protein expression for identifying surrogate markers for colon cancer progression, (b) a non-2D gel approach for dissecting complex mixtures of membrane proteins, (c) proteins that inhibit cytokine signal transduction, (d) proteins that are involved in the intricate pathway that leads to programmed cell death (apoptosis).

4. Human Kidney Glomerulus Proteome and proposition of a method for native protein profiling
Akira Tsugita, K. Miyazaki, Y. Yoshida and T. Yamamoto, NEC Proteomics Reseach Center and Niigata Univ. Medical Faculty, Japan

To elucidate molecular mechanism of a chronic nephritis, the following proteome research of kidney glomeruli has been initiated. Pieces of cortex of kidney with normal appearance were obtained from patients underwent surgical nephrectomy due to renal tumor. Glumeruli preparation were carried out from the cortex by a standard sieving process using four sieves. The glomeruli on the 150 μm sieve were collected and further purified by picking up under a phase-contract microscopy. The glomeruli were spun down, homogenized in 2-DE lysis buffer and incubated.

2-DE was carried out from the glomeruli preparation in the standard method (25×20 cm) and about 1500 protein spots were separated. Identification of protein has been carried out by N-and-C-terminal sequencings and peptide mass fingerprinting with MALDI-TOF-MAS. 200 spots have been identified.

Besides, a new method has been developed to obtain native protein profiling. The first dimension is in liquid phase on an isoelectric chromato-focusing column and the second dimension is by non-polar chromatography and molecular sieving chromatography or a special designed reverse-phase chromatography.

Track III-D-2: Genetic Data Issues

Chair: H. Sugawara

1. Genetic diversity in food legumes of Pakistan as revealed through characterization, evaluation and biochemical markers

Abdul Ghafoor and Asif Javaid, Plant Genetic Resources Institute, National Agricultural Research Center, Islamabad, Pakistan

Pakistan enjoys four distinguish seasons a year that enables to produce winter as well as summer legumes. Winter legumes consists of Chickpea (Cicer arietinum L.), lentils (Lens culinaris), peas (Pisum sativum), grass pea (Lathyrus sativus) and faba bean (Vicia faba), whereas summer legumes are mungbean (Vigna radiata), black gram (Vigna mungo), cowpea (Vigna unguiculata) and moth bean (Vigna oconotifolium). Common bean (Phaseolus vulgaris) is confined to high mountainous region of northern areas ranging the altitude 1000 to 2400 masl. These legumes have been collected and preserved in the gene bank for short duration (5-10 years) at 4 °C, medium term (15-20 years) at 0 °C and long term (more than 50 years) at -20 °C. The number preserved in the gene bank is 2065 (chickpea), 805 (lentil), 104 (peas), 100 (lathyrus), 101 (faba bean), 626 (mungbean), 646 (black gram), 199 (cowpea), 85 (moth bean) and 101 (common bean). About 80% of this germplasm has been characterized and evaluated for quantitative traits. Forty accessions of wild chickpea and one wild Vigna spp. have also been preserved.

The germplasm of black gram (250 accessions), mungbean (60 accessions), lentil (350 accessions), chickpea (350 accessions), wild chickpea (40 accessions), peas (104 accessions), cowpea (173 accessions) and wild Vigna spp. (one accession) have been evaluated for SDS-PAGE and except peas and wild chickpea, a low level of genetic diversity was observed for all the material evaluated. This situation lead to use of DNA markers, therefore 40 accessions of black gram and ten accessions of lentil were used for RAPD analysis that gave higher level of genetic diversity than SDS-PAGE. It was concluded that legume genetic resources should be characterised and evaluated along with biochemical analyses including protein and DNA markers for better gene bank management. This comprehensive data will lead to establishment of core collections. Either of the legumes mentioned above are mandate crop of one or other international centres except black gram and moth bean, although later is less important. Black gram has been identified a potential crop for most of Asian countries including India, Nepal, Bangladesh, Sri Lanka, Pakistan, Philippines, Thailand, Korea, Japan, Taiwan, China, etc. It is also recognized as an important crop in a part of African continent. Low genetic diversity coupled with low stability is a characteristic of this crop that could be minimized by developing a sound linkage between black gram growing countries and PGRI could serve as a regional gene bank for black gram preservation, evaluation and distribution of germplasm.

2. Relatedness of tRNA Sequences to Class I/II Aminoacyl-tRNA Synthetase: Families Detected by a Mathematical Logical Approach

Éena Jakó and Péter Ittzés, Department of Plant Taxonomy and Ecology, Eötvös Lorand University, Budapest, Hungary

Transfer RNA sequences from databases were analyzed for sequence features correlated with known classes of aminoacyl-tRNA synthetase enzymes. Previous analyses, based on statistical approaches did not find nucleotides predictive of synthetase class membership. In the present study, besides of conserved sets of nucleotides that contribute to the positive recognition, emphasis is made on detecting the sets of negative counterparts or anti-determinants that hinder the non-cognate aaRSs. The primary structures of macromolecules are viewed as some finite sets of linearly ordered symbols (nucleotide or amino acid bases), with some relations on them. The sequence information is rigorously defined by a system of logical functions that are considered as molecular descriptions. Consequently, whether any finite length, arbitrary set of symbols belong to a certain class of structures is assignable using relatively simple mathematical logical criterions. The biological adequacy of the method was tested on samples of different species, detecting relatedness of tRNA sequences to corresponding aminoacyl-tRNA synthetase families. The average proportion of correctly recognized sequences (including highly variable mtRNAs) was 93%. The proposed method may help to clarify the underlying logic of biochemical interactions between tRNAs and aaRSs. Furthermore, it can be used for detecting distant homologies in relatively short nucleotide or protein sequences, when application of statistical methods is not appropriate.

3. Visualization and Correction of Prokaryotic Taxonomy Using Techniques from Exploratory Data Analysis

T. G. Lilburn, American Type Culture Collection, USA
G. M. Garrity, Bergey's Manual Trust and Department of Microbiology and Molecular Genetics, Michigan State University, USA

There are, at present, over 5,700 named prokaryotic species. There has long been a need to organize these species within a comprehensive taxonomy that relates each species to all the others. For some years, researchers have been sequencing the small subunit ribosomal RNA genes of many prokaryotes, initially to try and establish the evolutionary relationships among all prokaryotes and subsequently in order to aid in the identification of prokaryotes both known and unknown. These sequences have become an almost universal feature in the description of new species. Thus, for the purposes of classification, the sequences are probably the most useful, universally described characteristic of the prokaryotes. Small subunit rRNA gene sequences were used by the staff of the Bergey's Manual Trust to establish prokaryotic taxonomy above the Family level only recently. This effort was facilitated by the application of techniques drawn from the field of exploratory data analysis to visualize the evolutionary relationships among large numbers of sequences and, hence, among the organisms they represent. We describe the techniques used to develop the first maps of sequence space and the techniques we are currently using to ease the placement of new organisms in the taxonomy and to uncover errors in the taxonomy or in sequence annotation. A key advantage of these techniques is that they allow us to see and use the complete data set of over 9,200 sequences. We also present plans for the development of a tool that will allow all interested researchers to participate in the maintenance and modification of the taxonomy.

4. Towards T-cell Epitope Design
Pandjassarame Kangueane, Meena K Sakharkar, Liew K. Meow, Nanyang Centre for Supercomputing and Visualisation, MPE, Nanyang Technological University, Singapore

Quantitative information on the types of inter-atomic interactions at the MHC-peptide interface will provide insights to backbone/sidechain atom preference during binding. Protein crystallographers have documented qualitative descriptions of such interactions in each complex. However, no comprehensive report is available to account for the common types of inter-atomic interactions in a set of MHC-peptide complexes characterized by MHC allele variation and peptide sequence diversity. The available x-ray crystallography data for MHC-peptide complexes in the Protein Databank (PDB) provides an opportunity to identify the prevalent types of inter-atomic interactions at the binding interface.

Two datasets, one consisting of 28 non-redundant class-I MHC-peptide complexes and another of 10 non-redundant class-II MHC-peptide complexes in the PDB were examined for inter-atomic interactions. Four types of such interactions namely - BB (backbone MHC - backbone peptide), SS (sidechain MHC - sidechain peptide), BS (backbone MHC - sidechain peptide) and SB (sidechain MHC - backbone peptide) characterize the MHC-peptide interface based on backbone and sidechain atom preference. We measured the percentage distribution of these interactions in a set of MHC-peptide complexes and identified the most common type among them.

We calculated the percentage distributions of four types of interactions at varying inter-atomic distances. The mean percentage distribution for these interactions and their standard deviation about the mean distribution is presented for each type. The prevalence of SS and SB interactions at the MHC-peptide interface is shown in this study. SB is clearly dominant at an inter-atomic distance of 3Å.

The prevalently dominant SB interaction at the interface suggests the importance of peptide backbone conformation during MHC-peptide binding. Currently available algorithms are well developed for protein side chain prediction upon fixed backbone templates. This study shows the preference of backbone atoms in MHC-peptide binding and hence emphasizes the need for accurate peptide backbone prediction in quantitative MHC-peptide binding calculations.

5. Intronless Genes in Eukaryotes
Meena Kishore Sakharkar and Pandjassarame Kangueane, Nanyang Technological University, Singapore

Eukaryotes have both intron-containing and intron-less genes and their proportion varies from species to species. Most eukaryotic genes are ‘‘multi exonic’’ with their gene structure being interrupted by introns. Introns account for a major proportion in many eukaryotic genomes. For example, the human genome is proposed to contain 24% introns and only 1.1% exons (Venter et al. 2001). Although most genes in eukaryotes contain introns, there are a substantial number of reports on intronless genes. We recently created a database (SEGE) for intronless genes in eukaryotes using GenBank 128 sequence data (http://intron.bic.nus.edu.sg/seg/). The eukaryotic subdivision files from GenBank were used to create a dataset containing entries that are reservedly considered as ‘‘single exonic’’ genes according to the ‘‘CDS’’ FEATURE convention. Single exon genes with prokaryotic architectures are of particular interest in gene evolution. Our analysis on this set of genes shows that structures are known for nearly 14% of their gene products. The characteristics and structural features of such proteins are discussed in this presentation.

Reference
Venter, C.J. et al. (2001) The sequence of the human genome. Science, 291, 1304-1351.

Track IV-A-2: Biodiversity II

Chair: Ji Liqiang, Institute of Zoology, Chinese Academy of Sciences, China

1. Shell Biodiversity Using Animation Technology
Sung-Soo Hong, Hoseo University, Korea
Bu-Young Ahn, Kye-Jun Lee and Ji-Young Kim, Bio-Resources Informatics Department, Korea

The world's natural history museums constitute an important storehouse of information about biodiversity. Although this information is regularly used for studies in systematic and natural history, its application to problems of importance to human well-being has been less frequent. Biodiversity is a new science that builds upon and combines the achievement at taxonomy, biology, biogeography, and ecology. It also draws on applied science such as conservation and natural resources management. A wide array at date types has been suggested as being relevant for biodiversity studies, ranging from molecular data to landuse data, early all of these data types can be structured around a core of 4 data elements : species, data, locality, and source, i.e. Theses data need to be digitized, cleaned-up, biogeography, and ecology. This paper is accompanied by a multimedia presentation of text, graphic, animation, virtual reality, and sound. This combination of data and its common visualization will provide a new insight about the interrelations among data. We developed a shell biodiversity using an animation technology (http://ruby.kisti.re.kr/~museumfs).

Cyber shell contents consists of five compartment including rare shells, marvelous shells, shell of the world, the shell of Korea and its story of shell. The database contains the pictures and related information of the shell. It implies not only animation display but also text information. The files of database were classified depending on the species, genus, family, order, and class and division of the shell. Pictures of shells are displayed and user may reach the image and virtual view information by clicking through the object displayed. This provides with various functions to multiplate, visualize and interact with image on the web. And every such transformations as translation, 360 degree rotation, and scaling can be applied in the picture interactively for the convenient and effective viewing. Information retrieval system using by corner transformation technique and multi-level grid file will be available for query search by future studies.

2. Building the Frog Contents System Using an Animation Technology
Sung-Soo Hong, Hoseo University, Korea
Bu-Young Ahn, Korea Institute of Science & Technology Information, Korea

In recent years, interest in surveying the biological resources of the country has increased greatly, with the goal of creating a national strategy to preserve biodiversity. Inventories and analyses of geographic, ecological, taxonomy and genetic diversity are key issues towards this goal. Frog dissection is mandatory part of biology or science courses offered in K-12 education and it is emphasized due to importance of the subject. Because of this, hundreds of thousands of frogs are dissected for the observation of their internal organs every year. This may not only result in environmental disruption but also has a risk of adversely affecting young students emotions as a side effect.

In the frog dissection system (http://ruby.kisti.re.kr/~museumfs), virtual dissection is enabled in order to eliminate these undesired effects and the factuality of organs is disguised using Photoshop to minimize the dislike of and aversion of students to the dissection process. In addition, the system was designed in such a way that, once a student replaces the dissected organs after observation is done, a frog is reanimated and jumps around so that the student does not treat the subject without care but instead treats it with respect for its life.

3. Biodiversity of Autotrophic Cryptogams in Antarctica
Asif Javaid, Abdul Ghafoor and Rashid Anwar, Plant Genetic Resources Institute, National Agricultural Research Center, Islamabad, Pakistan

Antarctica, the southernmost continent is a landmass of around 1.36 million square kilometers 98 percent covered by ice up to 4.7 kilometers thick. The continent remained neglected for decades after discovery, scientific research was initiated in early 1940s. Two species of phanerogams have been reported, whereas most of studies are carried out on cryptogams like algae, lichens and bryophytes. There are 700 species of terrestrial and aquatic algae in Antarctica, 250 lichens and 130 species of bryophytes including100 species of mosses and 25-30 species of liverworts. The species composition and abundance are controlled by many environmental variables, such as nutrients, availability of water and increased ultraviolet radiation resulting from the depletion of the ozone hole. These cryptogams can be found in almost all areas capable of supporting plant life in Antarctica and exhibit a number of adaptations to the Antarctic environment. There is a need to apply molecular and cellular techniques to study biodiversity and genetic characteristics of flora of this region. Biochemical techniques including DNA sequencing and microsatellite markers are being used to obtain information about the genetic structure of plant populations. These analyses are designed to assess levels of biodiversity and to provide information on the origin, evolutionary relationships and dispersal patterns. Flora of Antarctica needs to be genetically evaluated for the characters related to survival in that unique environment that can be incorporated into the economically important plants using transformation.

4. Automatic Mapping and Monitoring of Invasive Alien Plant Species, the South African Experience
J. M. K. Kandeh, J. L. Campos dos Santos and L. Kumar, International Institute for Geo-Information Science and Earth Observation, The Netherlands

Invasive alien plants are a huge problem in South Africa, affecting about 8.28% (10.1 million hectares of land) of the country. When converted to dense stands, this amounts to about 1.7 million hectares, and the problem is spreading rapidly. There is growing concern over the increasing rate at which the alien plants are replacing indigenous plants.

In response to the call of the convention on Biological diversity (UNEP, 1994), South Africa has over the years made efforts in compiling data on invasive alien plant species. A lot has been done in collating information on the distribution, abundance and habitat types of invasive alien plants, the role of biological agents in control of invasive alien plants, and modeling water use and spread of alien invasive plants.

Data on invasive alien plants in some part of the country are still weak and hence do not produce a comprehensive picture of alien plants invasion in the country. In the Greater St. Lucia Wetland Park of KwaZulu-Natal, the South African Government is implementing a mapping and control program on invasive alien plant species.

Control of invasive aliens species in the Wetland Park is also undertaken by a number of other organisations including private landowners, sugar cane farmers and forest plantation owners. There is lack of a standardised methodology with regards to data capture amongst the organisations. There are differences in data formats, map projections, little or no data exchange taking place, and most of the data on invasive alien plants held are not in computerized format. Consequently, there is very little information on the extent and distribution of invasive alien plants in the Greater St. Lucia Wetland Park.

This paper presents the development of a prototype geographic information system, which integrates data from various organisations in the Wetland Park. Integrating data from various organizations requires standardisation in data acquisition methodology, data representation and data management amongst the organisations.

In standardising data acquisition methodology, the methodology of Le Maitre and Versfed developed for mapping invasive alien plants at a 1:50,000 scale for a fynbos catchment management system was used, with the density classes grouped into four classes instead of seven without interfering with the class boundaries.

Using the Structured Systems Analysis Development Methodology, a prototype information system (APMIS) has been designed, tested and implemented. APMIS integrates data from various organisations in The Greater St. Lucia Wetland Park. APMIS is capable of providing geographic information on extent and distribution of invasive alien plants, assess eradication status of mapped areas, and provide operation maps of areas to be cleared. The APMIS strategy can be applied elsewhere where invasive alien plants are a problem and requires a coordinated approach both in mapping and control amongst all key players.

Keywords: Invasive Alien Plants, Geographic Information Systems, Biological Diversity, Systems Development Methodology

---

5. An Introduction of Chinese Biodiversity Information System

Ji Liqiang, Institute of Zoology, Chinese Academy of Sciences, China

Chinese Biodiversity Information System (CBIS) is a nation-wide distributed information system that collects, arranges, stores and disseminates data/information refers to biodiversity in China. It consists of a center system, 5 disciplinary divisions and dozens of data source. The Center System of CBIS is located in the Institute of Botany, Chinese Academy of Sciences, Beijing. The 5 divisions are Zoological Division (in Institute of Zoology, CAS, Beijing), Botanical Division (in Institute of Botany), Microbiological Division (in Institute of Microbiology, CAS, Beijing), Inland Wetland Biological Division (in Institute of Hydrobiology, CAS, Wuhan) and Marine Biological Division (in South China Sea Institute of Oceanology, CAS, Guangzhou). The data sources cover 15 institutes in CAS and includes botanical garden, field research station, museum, cell bank, seed bank, culture collection and research group. The Center System is response for building up and maintaining integrated and national-scale biodiversity database, environmental factor and vegetation database, model base and expert system in ecosystem level, and platform and tools of modeling and expert system. The Disciplinary Divisions are response for building up and maintaining database, model base and expert systems on their fields focused on data and information of species level. Data Sources are response for building up and maintaining database based on their local situation and disciplinary character, combining with GIS technology to present biodiversity information and data both in table and graphics.

82 databases have been set up in CBIS and been improved gradually, more than 590,000 records has been collected and inputted into CBIS database system, and most of them could be accessible from the Internet. They includes species inventory databases, endangered and protected species databases, ecosystem databases, specimen databases, botanical garden databases, culture collection database, cell bank database, economical species databases, etc.

In species inventory databases of animal, plant and microorganism, there are data of systematics, name, distribution, habitat and reference. In database of endangered and protected species, there are data of grade of protection, reason of endangered, measurement of protection, picture, etc. In database of specimen, there are data of collection, identification, storage and catalogue of species. CBIS recognizes the importance of metadata to data sharing and exchanging in its initial period and then sets up a series of standard of metadata in CBIS participating institutes. They include standard of metadata of dataset, data dictionary, metadata of institution and staff in CBIS. The metadata of dataset consists of 6 parts: information of dataset identity information, data collection, data management, data description, data accessing and metadata management. All databases of CBIS must be accompanied with a metadata file or table when they are put on the Internet or exchanged with other institutions.

6. Biodiversity Issues in Taiwan

Shang-Shyng Yang and Jong-Ching Su, National Committee for CODATA/Taiwan and Department of Agricultural Chemistry, National Taiwan University, Taiwan

In order to conserve and protect the very rich biological resources that have evolved in a unique natural environment, the government in Taiwan has set up a special committee and assigned a government agency, both at the cabinet level, to be in charge of planning and implementing relevant programs, respectively. Convening "Prospects of Biodiversity, Biodiversity-1999 and Biodiversity in the 21st Century" symposia has been the main means of building the national consensus to identify issues to be studied, which have motivated scientists to initiate the challenging task with the support of research funding from related agencies. There are 6 national parks, 18 nature reserves, 13 wildlife and 24 nature protection areas, totally covering 12.2% of the land area. The Policy Formulating Committee for Climate Changes has recommended the enforcement of education on biodiversity (including all levels of school and general public education), and formulated the working plans on the national biodiversity preservation and bioresources survey. The research programs in progress, supported by the national funding, include surveys on species, habitants, ecosystems and genetic diversities, long-term monitoring of diversity, sustainable bioresource utilization and compilation of flora of Taiwan. Increase in the number of scientific publications and increased emphasis placed by news media show the increased concern of both academic and public domains on biodiversity issue. Besides, the material and information databases related to the biological resources of various categories have been established and revised regularly. The following bioscience databases have been established in Taiwan: National plant genetic resources information system, Multimedia databank of Taiwan wildlife, Taiwan Agricultural Institute plant information system, Distribution and resources of fishes in Taiwan, Herbaria at many sites, Cell bank, Asian vegetable genetic resources and seeds, Database of pig production, Registry of pure-bred swine, Mating, furrowing, performance and transfer of ownership of pure-bred swine, Food marketing information system database, Food composition table in Taiwan, Database on heavy metals in Taiwan soils, Greenhouse gases emission from agriculture, Global change database generated in Taiwan.

Keywords: Biodiversity, national park, public education, bioscience, conservation policy, database

Track IV-B-2: Bioinformatics

Chair: Takashi Kunisawa, Science University of Tokyo, Japan

Biologists are facing the challenge of organizing and integrating a vast amount of data and information, which are mainly produced by genome projects. This session focuses primarily on quality controls in sequence databases. Phylogenetic analyses of sequence data are also included in the scope.

1. Unweaving regulatory networks: automated extraction from literature and statistical analysis
Andrey Rzhetsky, Columbia Genome Center, Columbia University, USA

In the first part of the talk I will describe our on-going effort to build a natural language processing system extracting information on interactions between genes and proteins from research articles. In the second part of the talk I will introduce an algorithm for predicting molecular networks from sequence data and stochastic models of birth of scale-free networks.

2. Genome rearrangements in the clinic and in evolution
David Sankoff, Centre de recherches mathematiques, Universite de Montreal, Canada

We analyze data on rearrangement breakpoints resulting from individual real-time cytogenetic events in order to help understand the distribution of multiple breakpoints in comparative maps. We compare breakpoint positions from four different databases, on reciprocal translocations, inversions and deletions in neoplasms, reciprocal translocations and inversions in families carrying rearrangements and the human-mouse comparative map. For each set of positions we construct breakpoint distributions for as many as possible of the the 44 autosomal arms.

We identify and interpret four main types of distribution:

1. The uniform distribution associated both with families carrying translocations or inversions, and with the comparative map,
2. Telomerically skewed distributions of translocations or inversions detected consequent to births with malformations,
3. Medially clustered distributions of translocation and deletion breakpoints in tumor karyotypes,
4. Bimodal translocation breakpoint distributions for chromosome arms containing telomeric proto-oncogenes.

3. PIR Integrated Databases And Data-Mining Tools For Genomic And Proteomic Research
Zhang-Zhi Hu, Winona C. Barker and Cathy H. Wu, Protein Information Resource, National Biomedical Research Foundation, Georgetown University Medical Center, Washington, DC, USA

The human genome project has revolutionized the practice of biology and the future potential of medicine. With the accelerated accumulation of high-throughput genomic and proteomic data, computational approaches are increasingly important for deriving scientific knowledge and hypotheses.

As an integrated public resource of protein informatics, the Protein Information Resource (PIR) provides many databases and analytical tools to support genomic and proteomic research and scientific discovery. The Protein

Sequence Database (PSD) is the major annotated protein database in the public domain, containing about 280,000 sequences covering the entire taxonomic range. To provide high quality annotation and promote database interoperability, the PIR uses rule-based and classification-driven procedures based on controlled vocabulary and accepted ontologies, and includes evidence attribution to distinguish experimentally determined from predicted protein features. PIR-NREF, a non-redundant database containing almost 1,000,000 proteins from PIR-PSD, Swiss-Prot, TrEMBL, GenPept, RefSeq, and PDB, provides a timely and comprehensive sequence collection with source attribution for protein identification, ontology development of protein names, and detection of annotation errors. The composite protein names in NREF, including synonyms and alternate names, and the bibliographic information from all underlying databases provide an invaluable knowledgebase for application of natural language processing or computational linguistics techniques to develop a protein name ontology. The iProClass database addresses the database interoperability issues arising from voluminous, heterogeneous, and distributed data. It provides comprehensive family relationships and functional and structural features for about 800,000 proteins in PIR-PSD, Swiss-Prot, and TrEMBL, with rich links to over 50 databases of protein families, functions, pathways, protein-protein interactions, post-translational modifications, structures, genomes, ontologies, literature, and taxonomy. The PIR databases are implemented in an object-relational database system and accessible online (http://pir.georgetown.edu) for exploration of proteins and their comparative analysis. It helps users to answer complex biological questions that may typically involve querying multiple sources and detect interesting relationships among protein sequences and groups.

The PIR is supported by the NIH grant P41 LM05798, iProClass is supported by the NSF grants DBI-9974855 and DBI-0138188, and the Protein Name Ontology project is supported by the NSF grant ITR-0205470.

---

4. Extraction of Phylogenetic Information from Gene Order Data
Takashi Kunisawa, Science University of Tokyo, Japan

Molecular phylogeny is frequently inferred from comparisons of nucleic or amino acid sequences of a single gene or protein family from different organisms. It is now known that there are a number of difficulties with this approach, for instance, correct alignment of sequence data, biased base (or amino acid) compositions among species, rate variation among sites and/or species, mutational saturation, and long-branch attraction artifact. Thus, development of new methods that can produce a reliable phylogenetic tree is an important issue. Here we present a simple method of reconstructing branching orders among genomes based on gene transpositions. We demonstrate that the occurrence or absence of a gene transposition event could provide empirical evidence for branching orders, being in contrast to the phenetic approaches of overall similarity or minimum distance. This approach is applied to evolutionary relationships among the completely sequenced Gram-positive bacteria. The complete genomic sequence data allow one to search for the target gene transpositions at a comprehensive level.

Earth and Environmental Data

Track I-C-3: Frameworks for Sharing Geographic Data

Chair: Santiago Borrero, GSDI Steering Committee, Colombia

This session reviews emerging technological and institutional models for widespread sharing of geographic data within and among large numbers of scientists and other users of geographic information. The frameworks described are complementary to each other. Individually and together they will facilitate expanded access and ease of use of geographic data across diverse and numerous scientific disciplines.

Among the framework initiatives to be addressed include:

1. The National Map,
2. Geospatial One-Stop,
3. The Geography Network, and
4. Frameworks for Sustainability of GIS Development in Low Income Countries.

From a U.S. perspective, the first three of these initiatives are all being developed within the standards and interoperability context of the U.S. National Spatial Data Infrastructure (NSDI). From a global perspective, these spatial database sharing efforts as well as those from many other nations are being developed within the context of the Global Spatial Data Infrastructure (GSDI) initiative.

---

1. Frameworks for Sustainability of GIS Development in Low Income Countries
Gilberto Camara, Director of Earth Observation, INPE, Brazil

This presentation discusses the development of Geographic Information System (GIS) software and technological approaches pursued in Brazil. Issues encountered in sustaining a complex technology in a large low income country (LIC) are outlined. In the process of describing the Brazilian experience, the prevalent assumption that LICs do not possess the complex technical and human resources required to develop and support GIS and similar technologies is challenged. Challenges, benefits and drawbacks of developing GIS software capabilities locally are examined and a number of important applications where local technology development has contributed to better understanding and cost-effective solutions are highlighted. Finally, some of the potential long-term benefits of a "learning-by-doing" approach and how other countries might benefit from the Brazilian experience are discussed.

2. The Geography Network
Clint Brown, ESRI, USA

Many now see the Internet as the most effective means of meeting the accelerating demand for geographically referenced information. Launched by ESRI in June, 2000, with the support of the National Geographic Society and many data publishers (EarthSat, GDT, WRI, US EPA, Tele Atlas, Space Imaging, etc.) the Geography Network <www.geographynetwork.com>, is a global collaborative and multi-participant network of geographic information users and providers including government agencies, commercial organizations, data publishers, and service providers, who use the Internet to share, publish, and use geographically referenced information. The Geography Network can be thought of as a large online library of distributed GIS information available to everyone. Users consult the Geography Network catalog, a searchable index of all information and services available to Geography Network users. A wide spectrum of simple to advanced GIS and visualization software technologies and online tools allow defining areas of interest, searching for specific geographic content, and can guide users to mapping services. Using any Internet browser, they access data that are physically located on servers around the globe, and can connect one or more sites at the same time. They can use digital map overlay and visualization, and combine and analyze many types of data from different sources. These data can be provided immediately to browsers or to desktop GIS software. Thousands of data layers are already available and Geography Network content is constantly increasing. Much of the content is accessible for free. Commercial content is also provided and maintained by its owners. Viewing or downloading of commercial content, or using commercial services, is charged in the Geography Network's e-commerce system. Becoming a provider is free and simple to do. The Geography Network uses open GIS standards and communication protocols, and serves as a test bed for data providers and the Open GIS Consortium. This presentation will show how the system works, explain the facilities provided, indicate the range of providers, describe the genesis of the system and its progress, and discuss future plans and directions.

3. Geospatial Information One-Stop
M. Robinson, Federal Geographic Data Committee, USA

The Geospatial One-Stop is part of a Presidential Initiative to improve effectiveness, efficiency, and customer service throughout the U.S. Federal Government. It builds upon the National Spatial Data Infrastructure (NSDI) and will accelerate its development and implementation. Geospatial One-Stop is classified as a Government-to-Government (G2G) project because it will focus on sharing and integrating Federal, State, local, and tribal data, and enable more effective management of government business. The vision is to spatially enable the delivery of government services.

The goals of Geospatial Information One Stop include providing fast, low-cost reliable access to Geospatial Data for government operations, facilitating G2G interactions needed for vertical missions such as Homeland Security, supporting the alignment of roles, responsibilities and resources, and establishing a methodology for obtaining multi-sector input for coordinating, developing and implementing geographic (data and service) information standards to create the consistency needed for interoperability and to stimulate market development of tools

The five major tasks identified in the Project Plan are: 1. Develop and implement data standards for NSDI Framework Data. 2. Fulfill and maintain an operational inventory (based on standardized documentation, using FGDC Metadata Standard) of NSDI Framework Data from Federal agencies, and publish the metadata records in the NSDI Clearinghouse network. 3. Publish metadata of planned acquisition and update activities for NSDI Framework Data from Federal agencies in the NSDI Clearinghouse network. 4. Prototype and deploy data access and web mapping services for NSDI Framework Data from Federal agencies. 5. Establish a comprehensive Federal portal to the resources described in the first four components (standards, priority data, planning information, and products and services), as a logical extension to the NSDI Clearinghouse network.

4. The National Map - Sharing Geospatial Data in the 21st Century
Barbara J. Ryan, U.S. Geological Survey, Reston, Virginia, USA

Over the last century, the United States has invested on the order of $1.6 billion and 33 million person hours in the standard (1:24,000 scale) topographic map series. These maps and associated digital data are the country's most extensive geospatial data infrastructure. They are also the only coast-to-coast, border-to-border coverage of our Nation's critical infrastructure - highways, bridges, dams, power plants, airports, etc. It is, however, an asset that is becoming increasingly outdated. These maps range in age from one year, those that were updated last year, to 57 years, those that have never been updated. The average age of these 55,000 maps is 23 years.

In January 2001, the Department of Interior's U.S. Geological Survey (USGS) undertook a decadal effort to transform the largely paper series to an online, seamless, integrated database known as The National Map. Extensive partnerships with local and State governments, other federal agencies, non-governmental organizations, universities and the private sector are being forged to construct The National Map. It is not a just a "federal" map, it is a "national" map -- an important distinction allowing greater leveraging of limited resources in order to fulfill the geospatial community's goal of "collect once, use many times."

These maps and related data touch, if not underpin, many sectors of the economy including the housing and development industry, agriculture, transportation, recreation, and emergency preparedness. After September 11th, the USGS provided more than 120,000 maps, hundreds of Landsat images and digital data files to assist with disaster planning, prevention, mitigation, and response efforts conducted at the local, State, and federal levels.

Coordination and standards-development mechanisms like the President's Geospatial One-Stop initiative, the Federal Geographic Data Committee, the Office of Management and Budget Circular A-16, and State-based geographic information consortia both advance and strengthen the policy framework for sharing geospatial data and other information assets of governments. The National Map, much like topographic maps in the last century, is a physical manifestation, in fact a visualization of this policy framework.

Track III-C-3: Earth and Environmental Data

Chair: Liu Chuang, Chinese Academy of Sciences, Beijing, China

1. Interactive Information System for Irrigation Management
Md Shahriar Pervez, International Water Management Institute, Sri Lanka
Mohammad Ahmadul Hoque, Surface Water Modelling Centre, Bangladesh

Irrigation management is a key to efficient and timely water distribution in canal command areas keeping in view the crop factors, and for irrigation management adequate and always updated information regarding the irrigation system is needed. This paper illustrates a GIS Tool for Irrigation Management which provides information interactively for decision making process. This Interactive Information System (IIS) has been developed to facilitate the operation and management of the command area development and to calculate the irrigation efficiency in the field level. At the basis of this development is geographic information systems (GIS) but gradually, this is being adapted to the kind of decision and management functions that lie at the heart of the planning process of any irrigation project. It also provides support to the design engineers to assess the impact of the design parameters of the System. This is an Arcview based GIS tool developed with the Avenue Codes by integrating the GIS and Relational Database Management System (RDMS). Effective integration of GIS with RDMS enhances performance evaluation and diagnostic analysis capabilities.

For this application real time topographic data are required which stored as spatially distributed datasets, back end RDMS has been used to store related attribute information, it lets an Irrigation manager to do some real time calculation and analysis which covers

a) Drawing of Detailed Canal and Drainage system on the basis of their category along with other spatial layers
b) Cross section profile of the canal
c) Comparison of cross sections
d) Long profile of the canal
e) Cut and Fill calculation of a Cross section in respect of the designed cross section of that particular section
f) Convince Calculation of a Particular Section
g) Calculate Area elevation curve for command area or any drawn area
h) Affected areas for the failure of any irrigation structure
I) Retrieval of current Irrigation Structure's Information along with image
j) Calculate the efficiency of the system.

Easy updating system of the associated database keeps the system always updated in respect of the real field situation. A very good user friendly Graphical User Interface at the front end helps the manager to operate the application easily. Using these "point on click" functions of this application an irrigation manager is capable to generate outputs in the form of Maps, Tables and Graphs which guide him to take prompt and appropriate decision with in few minutes.

2. Results of a Workshop on Scientific Data for Decision Making Toward Sustainable Development: Senegal River Basin Case Study

Paul F. Uhlir, U.S. National Committee for CODATA, National Research Council, USA
Abdoulaye Gaye, Senegalese National Committee for CODATA, Senegal
Julie Esanu, U.S. National Committee for CODATA, National Research Council, USA

Scientific databases relating to the environment, natural resources, and public health on the African continent are, for various reasons, difficult to create and manage effectively. Yet the creation of these and other types of databases-and their subsequent use to produce new information and knowledge for decision-makers-is essential to advancing scientific and technical progress in that region and to its sustainable development. The U.S. National Committee for CODATA collaborated with the Senegalese National CODATA Committee to convene a "Workshop on Scientific Data for Decision-Making Toward Sustainable Development: Senegal River Basin Case Study," which was be held on 11-15 March 2002, in Dakar, Senegal. The workshop examined multidisciplinary data sources and data handling in the West Africa region, using the Senegal River Basin as a case study, to determine how these data are or can be better used in decision making related to sustainable development. This presentation provides an overview of the workshop results and a summary of the published report.

3. Study on Spatial Databases of Chinese Ecosystems

Yue Yan-zhen, Chinese Academy of Sciences, China

The spatial databases construction of Chinese ecosystems is based on Chinese Ecosystem Research Network (CERN), Chinese Academy of Sciences (CAS). In order to meet the challenges of understanding and solving the issues of resources and environment at the regional or other larger scales, and with the support of Chinese Academy of Sciences, CERN started to be constructed in 1988. CERN consists of 35 ecological stations on agriculture, forest, grassland, lake and bay ecosystems, which produce a lot of data by monitoring and measurement every day. The quality of these data is control by 5 sub-centers of CERN, including water, soil, atmosphere, biological and aquatic sub-center. At last, all these enormous calibrated data including spatial data are collected in synthesis center.

We constructed the spatial databases to connect the enormous monitoring data with ecological spatial information. This study of the spatial databases includes:

1. Standard of spatial data classification
2. Structure of spatial databases
3. Functions of special databases
4. Management of special database
5. serving of net share
6. policy of data share

Key words: ecosystem network; Geographic Information System; Data Share

4. Development of the Global Map: National and Cross-National Coordination
Robert A. O'Neil, Natural Resources Canada, Ottawa, Canada

The Global Map is geospatial framework data of the Earth's land areas. This framework will be used to place environmental, economic and social data in its geographic context. The Global Map concept permits individual countries to determine how they will be represented in a global data base consisting of 8 layers of standardized data: administrative boundaries, drainage, transportation, population centres, elevation, land cover, land use and vegetation cover at a data density suitable for presentation at a scale of 1:1M. Usually it is the national mapping organizations that contribute data of their country to the Global Map, which is then made available at marginal or no cost.

At present, 94 nations have agreed to contribute information to the Global Map and an additional 42 are considering their participation. To date, coverage has been completed and is available for 11 countries.

While there is a wealth of source data available for this undertaking, not all nations have the capacity to evaluate the source data sets, make corrections and transform them into a contribution to the Global Map. A proposal to relax the specifications in order to hasten the completion of the Global Map will have to be balanced with the problems of dealing with heterogeneous databases, particularly in the integration, analysis and modeling.

Track III-D-4: The Use of Artificial Intelligence and Telematics in Environmental and Earth Sciences

Jacques-Octave Dubois, France and Alexei Gvishiani, Russia

New tools such as artificial intelligence algorithms are needed to effectively manage and process the vast amounts of environmental and earth science data.

Given that databases are increasingly widespread, telematics techniques (computer-based and telecommunications techniques) are needed to handle algorithms. In other words, to process this considerable amount of information, clustering algorithms must be adapted for and applied in computer networks.

The two books (Editions Codata, Springer) published by the two Co-chairs will be showcased at this session.

1. Application de l'Intelligence Artificielle et Télématique dans les Sciences de la Terre et de l'Environnement
Jacques-Octave Dubois, France
Alexei Gvishiani, Russia

Presentation of the the book : Artificial Intelligence and Dynamic Systems in Geophysical Applications. By A. Gvishiani and J.O. Dubois , Schmidt United Institute of Physics of the Earth RAS, CGDS and Institut de Physique du Globe de Paris.

This volume is the second of a two-volume series written by A. Gvishiani and J.O. Dubois.
The series presents the application of new artificial intelligence and dynamic systems techniques to geophysical data acqusition, management and studies. Most of the mathematical models, algorithms and tools presented were developed by the authors. The first volume of the series, published in 1998, is entitled "Dynamical Systems and Dynamic Classification Problems in Geophysical Applications." It is devoted to the application of dynamic systems, pattern recognition and finite vector classification with learning to a variety of geophysical problems.

The book "Artificial Intelligence" introduces geometrical clustering and fuzzy logic approaches to geophysical data analysis. A significant part of the volume is devoted to applying the artificial intelligence techniques introduced in volumes 1 and 2, to fields such as seismology, geodynamics, geoelectricity, geomagnetism, aeromagnetics, topography and bathymetry.

As in the first volume, this volume consists of two parts, describing complementary approaches to the analysis of natural systems. The first part, written by A. Gvishiani, deals with new ideas and methods in geometrical clustering and the fuzzy logic approach to geophysical data classification. It lays out the mathematical theory and formalized algorithms that form the basis for classification and clustering of the vector objects under consideration. It lays the foundation for the second part of this book which is the use of this classification in the study of dynamical systems.

The second part, written by J.O. Dubois, is concerned with various theoretical tools and their applications to modeling of natural systems using large geophysical data sets. Fractals and dynamic systems are used to analyse geomorphological (continental and marine), hydrological, bathymetrical, gravimetrical, seismological, geomagnetical and volcanological data.

In these applications chaos theory and the concept of self-organized criticality are used to describe the evolution of dynamic systems.

The first volume is devoted to the mathematical and algorithmical basis of the proposed artificial intelligence techniques; this volume presents a wide range of applications of those techniques to geophysical data processing and research problems. At the same time it presents a reader with another algorithmic approach based on fuzzy logic and geometrical illumination models.

Many readers will be interested in the two volumes (vol.1, J.O. Dubois, A. Gvishiani "Dynamic Systems and Dynamic Classification Problems in Geophysical Applications" and the present vol.2, A. Gvishiani, J.O. Dubois "Artificial Intelligence and Dynamic Systems in Geophysical Applications") as a package.

---

2. The Environmental Scenario Generator (ESG) a Distributed Environmental Data Mining Tool
Eric A. Kihn, NOAA/NGDC, Boulder, CO, USA
Dr. Mikhail Zhizhin, RAS/CGDS, Moscow, Russia

The Environmental Scenario Generator (ESG) is a network distributed software system designed to allow a user running a simulation to intelligently access distributed environmental data archives for inclusion and integration with model runs. The ESG is built to solve several key problems for the modeler. The first is to provide access to an intelligent ?data mining? tool so that key environmental data can not only be retrieved and visualized but in addition, user defined conditions can be searched for and discovered. As an example, a user modeling a hurricane?s landfall might want to model the result of an extreme rain event prior to the hurricane?s arrival. Without a tool such as ESG the simulation coordinator would be required to know:

- For my region what constitutes an extreme rain?
- How can I find an example in the real data of when such an event occurred?
- What about temporal or spatial variations to my scenario such as the finding the wettest week, month or year?

If we consider combining these questions across multiple parameters, such as temperature, pressure, wind speed, etc. and then add multiple regions and seasons the problem reveals itself to be quite daunting.

The second hurdle facing a modeler who wants to include real environmental effects in the simulation is how to manage many discreet data sources. Often simulation runs face tight time deadlines and lack the manpower necessary to retrieve data from across the network, reformat it for ingest, regrid or resample it to fit the simulation parameters, then incorporate it in model runs. Even if this could be accomplished what confidence can the modeler have in the different data sources and their applicability to the current simulation without becoming expert in each data type? The unfortunate side effect of this is that the environment is often forgotten in simulations or a single environmental database is created and ?canned? to be replayed again and again in the simulation.

The ESG solves this problem for the modeler by providing a 100% Java platform independent client with access to both data mining and database creation capabilities on a network distributed parallel computer cluster with the ability to perform fuzzy logic based searching on an global array of environmental parameters. By providing intelligent instantaneous access to real data it ensures that the modeler is able to include realistic, reliable and detailed environments in their simulation applications.

This demonstration will present the results of data-mining, visualization, and a domain integration tool developed in a network distributed fashion and applied to environmental modeling.

3. Satellite Imagery As a Multi-Disciplinary Tool for Environmental Applications
Herbert W. Kroehl, Eric A. Kihn, NOAA/NGDC, USA
Alexei Gvishiani, Mikhail N. Zhizhin, RAS/CGDS, Russia

Satellite technologies offer a unique opportunity to monitor the earth and its environment. Environmental satellite data, which initially focussed on "in situ" measurements of the ambient environment, are taking advantage of remote sensing technology through the use of imagers and sounders. Visible, infrared, microwave and ultraviolet emissions are now recorded across a swath as large as 3,000 km by instruments on operational meteorological and earth observing satellites. The resulting radiances are used to compute a disparate set of parameters serving very different scientific disciplines, e.g. space physics and sociology.

What environmental parameters are routinely computed from imagery and soundings recorded on satellites? The imagery on operational weather satellites are used to monitor clouds, snow, ice and solar activity and to construct profiles of atmospheric temperature, humidity and ozone. The same images were found to be useful in assessing the state of the environment, detecting wildfires, tracking the flow of ash from volcanoes, and assessing population dynamics. In addition to improvements for operational instruments, imagers on earth observing systems are used to assess environmental health, classify vegetation, to assess the effects of natural hazards, and to build digital elevation models.

But when the same data are used for many different applications, one scientist's signal becomes another scientist's noise, and it becomes important to classify different environmental signals contained in an image. In addition, data mining techniques need automatic classification of images, especially when these images are so voluminous.

A sample of the diverse use of images recorded on weather and earth observing satellites will be presented as a prelude to the need for mathematical techniques to classify information contained in the images.

4. Development of the Space Physics Interactive Data Resource- II (SPIDR II) Experiences Working in a Virtual Laboratory Environment
Eric A. Kihn, NOAA/NGDC, USA
Mikhail Zhizhin, RAS/CGDS, Russia
Alexei Gvishiani, RAS/CGDS, Russia
Herbert W. Kroehl, NOAA/NGDC, USA

SPIDR 2 is a distributed resource for accessing space physics data which was designed and constructed jointly at NGDC and CGDS to support requirements of the Global Observation and Information Network (GOIN) project. SPIDR is designed to allow users to search, browse, retrieve, and display Solar Terrestrial Physics (STP) and DMSP satellite digital data. SPIDR consists of a WWW interface, online data and information, and interactive display programs, advanced data mining and data retrieval programs.

The SPIDR system currently handles the following: DMSP visible, infrared and microwave browse imagery, ionospheric parameters, geomagnetic 1.0 minute and hourly value data, geophysical and solar indices, GOES x-ray, plasma, and magnetometer data, cosmic ray, solar radio telescope, satellite anomaly and city lights data sets. The goal is to manage and distribute all STP digital holdings through the SPIDR system providing comprehensive and authoritative on-line data services, analysis and numerical modeling to the space physics community.

The successful cooperation between NGDC and CGDS has produced the development of a SPIDR-I mirror in 1997, development and launch of SPIDR-II servers in Boulder, Moscow, and Sydney in 1999, additional SPIDR II mirrors in South Africa and Japan in 2000, and the development of a new satellite data systems prototype in 2001.

This presentation will present details of technologies, and methodologies that were successful in producing exceptional results from a geographically distributed team working in a virtual laboratory environment.

5. An Automatic Analysis of Long Geoelectromagnetic Time Series: Determination of the Volcanic Activity Precursors

J. Zlotnicki, Observatorie de Physique du Globe de Clermont-Ferrand, France
J-L. LeMouel, Director of the department of Geomagnetism, Institut de Physique du Globe de Paris, France
S.Agayan, Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia
Sh. Bogoutdinov, Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia
A. Gvishiani, Director of the Center of Geophysical Data Studies and Telematics Applications, IPE RAS, Russia
V.Mikhailov, Institute for the Physics of the Earth RAS, Russia
S.Tikhotsky, Institute for the Physics of the Earth RAS, Russia

The new methods developed for the geophysical long time series analysis, based on the fuzzy logic approach. These methods include the algorithms for the determination of anomalous signals. They are specially designed and very efficient in the problems where the definition of anomalous signal is fuzzy, i.e. the general signature, amplitude and frequency of the signal can not be prescribed a priory, as in the case of seeking for the precursors of natural disasters in geophysical records. The developed algorithms are able to determine the intervals of the record that are anomalous withrespect to the background signal presented at the record. Another part of algorithms deal with the morphology analysis of signals. These algorithms were applied for the analysis of the electromagnetic records over La Fournaise volcano (Reunion island). For several years five stations measured the the electric field along different directions. The signals specific for the eruption events are determined and correlated over several stations. Another types of signals that correspond to storms and other sources are also determined and classified. The software is designed that helps to analyze the spatial distribution of activity over stations.

6. Application of telematics approaches for solving the problems of distributed environmental monitoring

M. Zgurovsky, National Technical University of Ukraine, Kiev Polytechnic Institute

The results of research carried out at the Cybernetics Gloushkov Center of National Academy of Sciences of Ukraine are presented. A review of the advanced developments in the field of distributed environmental monitoring is given.

Among the presented developments - the interactive system of modeling and prognosis of ecological, economic and other processes on the basis of observations for support of taking up quick control decisions. The system is based on the inductive method of arguments group accounting used for automatic extraction of the substantial information from the measurement data. The efficiency of the system is demonstrated on applications of modeling and prognosis of dynamics changes of animal plankton concentration, number of microorganisms in contaminated soil and others.

The designs of the mobile laboratory of the quick radiation monitoring (RAMON) and of the automated system for research of subsoil water processes (NADRA) are presented. Problems of the user interface intelletualization in geophysical software are considered.

Track IV-B-5: Seismic Data Issues

Chair: A. Gvishiani, Director of the Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia

1. Clustering of Geophysical Data by New Fuzzy Logic Based Algorithms

S.Agayan, Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia
Sh. Bogoutdinov, Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia
A. Gvishiani, Director of the Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia
M. Diament, Institut de Physique du Globe de Paris (IPGP), France
V.Mikhailov, Institute for the Physics of the Earth RAS, Russia
C. Widiwijayanti, Institut de Physique du Globe de Paris (IPGP), France

A new system of clusterization algorithms, based on geometrical model of illumination in the finite-dimensional space, has been developed recently, using fuzzy sets approach. The two major components of the system are RODIN and CRYSTAL algorithms. These two efficient clusterization tools will be presented along with their applications to seismological, gravity and geomagnetic data analysis. The regions of Malucca Sea (Indonesia) and Gulf of San Malo (France) are under consideration. In the course of study of the very complicated geodynamics of the Malucca sea region the clusterization of earthquakes hypocenters with respect to their position, type of faulting and horizontal displacement strike was performed. The results of this procedure made more clear the stress pattern and hence the geodynamical structure of the region. RODIN algorithm was also applied for clustering of the results of anomalous gravity field pseudo-inversion over this region. It improved the solution considerably and helped to determine the depths and horizontal positions of sources of the gravity anomalies. The obtained results correlate well with the results of the local seismic tomography and gravity inversion. In the region of Gulf of San Malo the developed algorithms was successfully used to investigate the structure of quasi-linear magnetic anomalies onshore and offshore.

2. Artificial Intelligence Methods in the Analysis of Large Geophysical Data Bases

A. Gvishiani, Director of the Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia
J.Bonnin, Institut de Physique du Globe de Strasbourg, France

The presentation is devoted to the different kinds of Artificial Intelligence Algorithms, oriented towards geophysical applications: syntactic pattern recognition, geometrical cluster analysis, time series processing and classification, dynamic pattern recognition with learning and other considerations. A big deal of the presentation is devoted to fuzzy logic and fuzzy mathematics applications to artificial intelligence algorithms development. The following geophysical and environmental applications will be presented: recognition of strong earthquake-proun areas in Alps-Perineas and Caucasus, syntactic classification of seismograms and strong ground motion records, identification of anomalies on geoelectrical and gravity data, use of clustering for the interpretation of geomagnetic data.

3. Geo- Environmental Assessment of Flash Flood Hazard of the Safaga Terrain, Egypt, Using Remote Sensing Imagary

Maged L. El Rakaiby, Nuclear Materials Authority, Egypt
Mohamed N. Hegazy, National Authority for Remote Sensing and Space Sciences, Egypt
Menas Kafatos, Center for Earth Observing and Space Research, GMU, USA

We emphasize the use of space images for detecting, interpreting and mapping elements of the geological and geomorphologic environment of the Safaga terrain, Egypt to monitor the geomorphologic elements causing flash floods. Safaga town and associated highways are highly affected by flash floods more than once every year. Information interpreted from space images is very useful for reducing flash flood hazard and adjusting the use of the Safaga terrain.

4. On the Modeling of Fast Variations of the Mode of Deformation of Lithospheric Plates

M. Diament, Organization Institut de Physique du Globe de Paris (IPGP), France
Dubois J.-O., Organization Institut de Physique du Globe (IPGP), France
Kedrov E., Center of Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia
M. Kovalenko, The State Research Institute of Aviation Systems, Russia
Mikhailov V., Institute for the Physics of the Earth RAS, Russia
Murakami Yu., Geological Siurvey of Japan, Japan

This paper discuses possible applications of the new recently obtainedexact solutions of the elasticity theory problems for the domains having corner points. Analysis of the solutions obtained demonstrated that the mode of deformation in the narrow zones along the boundary of such bodies close to the corner points strongly depends on the work of the surface forces released in these points.

Exact solutions for a rectangle principally differs from the classical exact solutions for unbounded domains (e.g. wedge, infinite stripe etc.) or for domains limited by smooth boundary. The explanation is in the fact that properties of corner points differs considerably from properties of the domain they belong to. In particular, such fundamental notion as surfent of area can not be introduced at the corner point, thus effect of this point can be calculated only as an additional work released in the corner point by some fictitious forces and/or torque which are additional to acting surface forces.

When some interval of a body boundary is of a high curvature or contains a corner point and when boundary loading does not neglect there, then small variations of the shape of the boundary in the vicinity of such interval or corner point can cause finite or even infinite variations of the specific energy. This actually means that Saint-Venant principle is not valid for the areas containing corner points. When boundary of an area is strongly irregular then solution depends on how boundary loading accommodates to the intervals of high boundary curvature.

The results obtained makes it possible to consider corner points of lithospheric plates as singular or "trigger" points, probably responsible for the fast observable changes of the mode of deformation along plate boundaries. These fast changes at the plate boundaries could arise not only in result of variation of boundary forces in the vicinity of corner points but also in result of changes of inner structure and/or rheology inside the plates. The last changes could arise as from decompaction of rocks in the vicinity of corner points (as a consequence of earthquakes, tectonic or thermal processes) or, vice versa; arise from rock compaction, taking place during periods of seismic quietness.

This investigation has been performed by the stuff of virtual laboratory on new solution of the elasticity theory designed and maintained by scientists from Russia, Japan, France and USA in the frameworks of joint project supported by International Science and Technology Center. Website designed supports teleconferences, exchange and presentation of results.

5. New Mathematical Approach to Seismotectonic Data Studies
M. Kovalenko and N. Tsybin, State Research Institute of Aviation Systems
Yu. Rebetsky, Institute of Physics of the Earth, Russia
Yu. Murakami, Geological Survey of Japan, Japan

The paper discuses possible applications of the new recently obtained exact solutions of the some classical problems of the elasticity theory for domains having ruptures. Analysis of the solutions obtained demonstrated that the solution for domains with ruptures is non unique. The explanation is in the fact that the properties of apexes of crack differs considerably from properties of the domain they belong to. The stress distribution strongly depends on the work of the surface forces released in these points. Practically it is a question of the work released on micro level. Thus effect of apexes of crack can be calculated only as an additional work released there.

The results obtained makes it possible to consider apexes of fault of lithospheric plates as trigger points, probably responsible for the fast observable changes of the mode of deformation. These fast changes could arise not only in result of variation of boundary forces in the vicinity of apexes of fault but also in result of changes of inner structure and/or rheology inside the plates. Actually it means, that the crack energy may change without increase/decrease of length of crack.

This study has been performed using virtual laboratory approach designed and maintained by scientists from Russia, Japan, France and USA in the frameworks of joint project supported by International Science and Technology Center. Web-site designed supports teleconferences, exchange and presentation of results.

Track IV-A-6:
Application 2D et 3D de systèmes SIG. Transopérabilité de gestion intégrée de bases à composantes cartographiques (2D and 3D Applications of GIS Systems: Interoperability of Integrated Cartographic Database Management)

Chairs: Jacques Segoufin, Institut de Physique du Globe, Paris, France et
Michail Zgurovsky, National Technical University of Ukraine, Kiev Polytechnic Institute, Ukraine

Selon les pays on se base, pour assurer la projection des éléments de carte sur le plan des données de surface courbe de notre planète, sur le choix de systèmes de références.

Le choix d'un ellipsoïde de référence est variable et diverses recommandations sont disponibles. De la nature des types de projection dépendent également la qualité des correspondances entre cartes mondiales, régionales et locales.

La situation des données acquises à jour sont de plus en plus effectuée dans le système UTM84.

Quelques projets européens visent à assurer les passages et de transferts plus faciles des " données techniques " locales pour les retraduire dans UTM84.

Pour la présente session il est proposé d'aborder en particulier les problèmes suivants :

- Aspect théoriques, problèmes de référentiels et de projection
- Réajustement des données issues de grilles différentes
- Intégration multiparamètres
- Organisation en réseaux des éléments d'un SIG
- Liaison des domaines continentaux océaniques
- Etats des grands projets internationaux (UNESCO, IGN, Geological Surveys)

1. Application of methods of space-distributed systems modeling in ecology
M. Zgurovsky, National Technical University of Ukraine, Kiev Polytechnic Institute

A review of the studies carried out at NTUU"KPI" and the Institute of Cybernetics of National Academy of Sciences of Ukraine is presented. Two-dimensional and three-dimensional equations of diffusion and heat - mass transfer are used as mathematical models. The models make it possible to take account of space distribution, structural non-uniformity and anomaly properties of physical processes of harmful impurities spreading in the atmosphere, open water pools and subsoil waters.

The considered processes are characterized by substantial distribution in space. Therefore, efficient methods of numerical solution of two- and three-dimensional model equations are presented.

The complexes of programs allowing to solve efficiently the problems of modeling, prognosis and estimation of ecological processes in various environment are given.

2. Une mission géographique et ethnopharmacologique sur les plantes toxiques de l'Ile Maurice
A. Fakim-Gurib, Université de l'Ile Maurice
P. van Brandt, Université catholique de Louvain Bruxelles Belgique

La région sud ouest de l'Océan Indien est une zone géographique privilégiée pour sa diversité biologique et elle est bien connue pour sa flore d'espèces endémiques. A l'île Maurice l'usage traditionnel est très commun mais beaucoup de plantes utilisées peuvent apporter des risques potentiels pour la santé.

Bien que certaines plantes soient connues pour contenir un grand nombre de composés biologiquement actifs aux nombreux effets bénéfiques pour l'homme et les animaux, certains de ces mêmes éléments devraient être sujet à des dosages car à cause de leur utilisation abusive, ils se sont avérés extrêmement toxiques et provoquent ainsi des effets néfastes à la santé. L'apparition de ces effets négatifs peuvent être très soudains ou prendre du temps pour se développer. Heureusement, il n'y a que relativement peu de plantes qui, lorsqu'elles sont ingérées causent des troubles dangereux pour l'organisme. Cependant des précautions préliminaires doivent être prises pour éviter des empoisonnements, en particulier chez les jeunes enfants. Par conséquent, une mission a été menée à l'île Maurice pour identifier ces plantes toxiques. Soixante neuf espèces ont été inventoriées comme étant potentiellement toxiques et comprennent des espèces depuis : Thevetia peruviana (Apocynacaeae) considérée comme extrêmement toxique, jusqu'à Dieffenbachia seguine (Araceae) considérée comme modérément à faiblement toxique. Il est à remarquer que deux plantes indigènes endémiques de la région sont aussi considérées comme ayant des propriétés toxiques : Cnestis glabra (Connaraceae) et Agauria salicifolia (Ericaceae).

Les résultats de cette mission illustrent les degrés différents de toxicité, les composants chimiques et leur effets.

Les effets bénéfiques à long terme de ces toxines ne doivent pas être sous estimés si l'on prend l'exemple de Taxus brevifolia qui a donné entre autre naissance au fameux Taxol.

Un autre aspect, qui mérite d'être pris en compte est le fait que le climat et les facteurs environnementaux ont une influence directe sur la phytochimie de la bio diversité florale locale.

3. Carte structurale de l'océan indien
J. Segoufin, Institut de Physique du Globe de Paris, France

Dans le cadre des activités de la CCGM (Commission de la Carte Géologique du Monde), sous la supervision de l'UNESCO, il a été décidé de créer un certain nombre de cartes géologiques, tectoniques, structurales englobant le domaine maritime pour lequel il y a maintenant beaucoup d'informations, la Commission pour la cartographie des fonds sous marins étant en charge de ce dernier domaine;

C'est ainsi, qu'il y a deux ans, il a été décidé d'éditer une carte structurale de l'océan indien.

Cette carte a pour but de faire la synthèse des connaissances sur cet océan, d'en montrer sa formation et son évolution à partir des données géophysiques recueillies par différents instituts. Cette carte a un but pédagogique de diffusion des connaissances et doit être diffusée dans les lycées, Collèges et Universités.

Après plusieurs essais, les limites géographiques de la carte ont été fixées de 0° à 155° E et de -71° S à 30° N. La carte sera éditée dans le système de projection de Mercator à une échelle de 1/10.000.000 ème

La carte structurale de l'océan indien est constituée de 4 feuilles :

Feuille1 :	0°	80° E
	-30° S	30° N
Feuille2 :	80° E	155° E
	-30° S	30° N
Feuille3	0°	80° E
	-71° S	-30° S
Feuille4	80° E	155° E
	-71° S	-30° S

L'ensemble des donnée qui sont accessibles actuellement figurera sur cette carte :
les courbes bathymétriques, les anomalies magnétiques, l'âge de la croûte océanique, tiré des anomalies magnétiques, les épicentres des seismes, divisés en deux classes : magnitudes supérieures à 6 et magnitudes inférieures à 6, les failles transformantes et zones de fracture, l'épaisseur sédimentaire, les zones de subduction, les axes de dorsale, les volcans actifs, les astroblèmes, les sites de forages DSDP et ODP ayant atteints la croûte océanique, les monts et plateaux sous-marins, etc…

En complément de cette carte, il a été également décidé de sortir une feuille physiographique, calculée à partir de la grille de Sandwell et al. et couvrant l'ensemble de l'océan indien en une seule feuille.

La plus grande partie des difficultés rencontrées dans la création de la carte structurale, consiste à rendre cohérentes, dans un même format des données qui proviennent d'horizons divers, nécessistant la plupart du temps un traitement préalable.

Actuellement les feuilles 1, 2, 3 sont terminées, la feuille 4 est en cours.

Une présentation de l'ensemble de la carte est prévue à la réunion de l'EUG à Nice en Avril 2003.

Le but de ce travail est de diffuser l'état actuel des connaissances sur l'océan indien grâce à cette série de cartes, mais également d'un support informatique interactif (CDROM), sur lequel les différentes informations apparaîtront sous forme de couches superposées pouvant être ajoutées ou supprimées à la demande.

La fin de ces travaux est programmé pour 2004, date à laquelle les versions papier et digitale de la Carte Structurale de l'Océan Indien seront présentées à la réunion de l'IGC à Florence.

---

4. Passerelle d'information sur les collections, spécimens et observations biologiques (ICSOB)
Guy Baillargeon, Agriculture and Agri-Food Canada

La Passerelle ICSOB est un prototype de moteur de recherche et de cartographie spécialisé sur les données d'observation et les spécimens biologiques des collections d'histoire naturelle. ICSOB répertorie les données disponibles par l'intermédiaire de réseaux de biodiversité accessibles sur l'Internet par voie de requêtes distribuées tels que l'Analyste d'espèces (TSA), le Réseau mondial d'information sur la biodiversité (REMIB) ou le Réseau européen d'information sur les spécimens d'histoire naturelle (ENHSIN). De façon analogue aux moteurs de recherche (tels que Google ou Altavista) qui aident à localiser des documents hypertextes, ICSOB récolte des noms dans les collections distribuées sur les réseaux de l'Internet et connecte les usagers directement aux sources de données originales. Les enregistrements de données transitent directement des gestionnaires autorisés de données primaires aux usagers finaux en temps réel. En outre, les enregistrements pourvus de coordonnées géographiques (longitude, latitude) sont reportés dynamiquement sur une carte du monde dont chacun des points de distribution est directement relié aux données originales. La Passerelle ICSOB fourni un point d'accès à des millions d'enregistrements individuels en provenance de plusieurs réseaux de biodiversité distincts. ICSOB est pleinement intégré à la version multilingue du Système d'information taxonomique intégré (SITI) facilitant l'accès aux données soit par l'intermédiaire de noms communs, de noms scientifiques ou de synonymes.

Medical and Health Data

Track I-D-2: The US National Library of Medicine's Visible Human Project® Data Sets

Chair: Michael J. Ackerman, National Library of Medicine, National Institutes of Health, USA

In the mid-1990s, the US National Library of Medicine sponsored the acquisition and development of the Visible Human Project® database. This image database contains anatomical cross-sectional images, which allow the reconstruction of three-dimensional male and female anatomy to an accuracy of less than 1.0 mm. The male anatomy is contained in a 15 gigabyte database, the female in a 40 gigabyte database.

This session will consist of four papers. The first will summarize the history of the Visible Human Project® and the development of the Visible Human data sets. We will then explore the problems encountered in the real-time navigation of such large image databases. The third paper will discuss the extraction of data from such a database, and the final paper will cover the problems of validation.

---

1. The Visible Human Project® Image Data Sets
From Inception to Completion and Beyond
Richard A. Banvard, National Library of Medicine, National Institutes of Health, USA

The Visible Human Project® Data Sets resulted from a recommendation of the National Library of Medicine (NLM) Board of Regents' 1987 Long Range Plan that stated the NLM should "thoroughly and systematically investigate the technical requirements and feasibility of instituting a biomedical images library." At the suggestion of an expert panel convened by the Board and reporting in April 1990 that - "NLM should undertake a first project, building a digital image library of volumetric data representing a complete normal adult human male and female. This 'Visible Human' project would include digital images derived from computerized tomography, magnetic resonance imaging, and photographic images from cryosectioning of cadavers." - the University of Colorado was contracted in August 1991 by NLM to undertake collection of this "Visible Human" image data set. In November 1994 the Visible Human Male data set was announced and released to the public, followed one year later by the Visible Human Female. The data sets are available via FTP at no cost, to anyone holding a no cost license. Each image: CT, MRI and cryosection is stored as a separate file; can be downloaded singularly or in any number up to the entire data set. Several mirror sites have been established to facilitate download for international license holders. The images also can be purchased on tape for a fee from the National Technical Information Services (NTIS). This session will include a discussion of the genesis of the Visible Human Project®, a description of the University of Colorado's cryosectioning procedures, and descriptions of several of the more interesting and notable outcomes developed by license holders who have used The Visible Human Project® Data Sets.

2. Visible Human Explorer
Hao Le, Flashback Imaging Inc., Canada
Brian Wannamaker, Sea Scan International Inc., Canada

The technology for imaging in medical applications continues apace. This increases the potential for improvements in medical research, diagnostic procedures, and patient care. On the other hand, the increase in imaging activity also increases the shear volume of data that must be dealt with. The imagery may be reviewed for immediate diagnostic procedures and discarded. Or it may be stored or archived for further use. However, storage or archiving is effectively discarding unless effective means for recovering the data exist. Accessibility is an essential component of developing and distributing new knowledge from growing data volumes. This paper will discuss specific approaches to improving accessibility of large image databases like that of the Visible Human Project. Real time navigation in 2D and 3D of image databases as well as user interfaces designed for public and academic use will be outlined. The presentation will be illustrated with some thousands of images from the Visible Human Project.

3. The NLM Insight Registration and Segmentation Toolkit
William Lorensen, GE Research, USA

In 1999, the National Library of Medicine (NLM) awarded six contracts to develop a registration and segmentation toolkit. The overall objective of the project is to produce an application programming interface (API) implemented within a public domain toolkit. The NLM Segmentation and Registration Toolkit supports image analysis research in segmentation, classification and deformable registration of medical images. This toolkit meets the following critical technical requirements identified by the National Library of Medicine:

- Work with the Visual Human Male and Female data sets.
- Provide a foundation for future medical image understanding research.
- Become a self sustaining code development effort.
- Accommodate periodic and incremental modifications and additions.
- Accommodate expansion to parallel implementations.
- Accommodate large memory requirements.
- Support a variety of visualization and/or rendering platforms.

In addition to the technical challenges presented by these requirements, the selected team and subcontractors, had to work as a distributed group. The software development experience of the groups also varied. Some members had created software for a large community while others had only developed software for their local groups. The team defined a web centric software development process modeled after the Extreme Programming approach that relies on rapid and parallel requirements analysis, design, coding and testing. Communication through web based mailing lists and bug trackers was supplemented with conventional telephone conferences.

The first public version of the software is scheduled for release in October, 2002.

This talk discusses the chronology of the project, the core architecture and algorithms as well as the light weight software engineering processes used throughout the project. Finally, we present lessons learned that will be of value to future distributed software development projects.

4. The Visible Human Data Sets: A Protoype and a Roadmap for Navigating Medical Imaging Data
Peter Ratiu, Harvard Medical School, Brigham and Women's Hospital, USA

The Visible Human Data Sets are to date the most complete, multi-modality data sets of human anatomy. The computational challenges posed have been widely discussed and many of them have been or are being solved by experts in various aspects of medical image analysis and medical informatics. Their approach, which has proved profitable, is to regard the Visible Human as a vast collection of bits, single and multi-channell images, with little regard to its intrinsic content B human anatomy. This approach allowed them to solve computational problems that had appeared overwhelming at the inception of the project: powerful servers can make available the individual images, manipulate and display them in various ways, on the desktop of end-users. An example of such solution implemented by computer scientists is the EPFL Server.

The more specific problem, of how to use this unique information in medical research has been less often addressed. One reason for this, is that the data is vast, its manipulation seemingly unwieldy for anatomists, until now more versed in using scalpels than mouse buttons. Another reason is, the inherent novelty of the data: for the first time, it opens the possibility of a quantitative approach to anatomy. However, this quantitative approach can be best exploited by first defining problems in anatomy, anthropology, pathology in these terms.

I will discuss two basic aspects of the Visible Human Project as a landmark data set:

1. The problem of establishing a universal anatomical coordinate system, with applications in basic research as well as clinical medicine (radiology, clinical imaging), and how the VHP can contribute to the solution.
2. The need for a quantitative comparative anatomy, as this is becoming apparent in a broad array of disciplines, ranging from physical anthropology to gynecology. I will present how the VHP data can be employed as a roadmap for navigating diverse data.

The aim of this presentation is to present the problems related to medical imaging data to experts in other fileds, in such manner, that it may spark a mutually profitable dialog with hitherto alien disciplines.

Track III-C-5: Emerging tools and techniques for data handling in developing countries

Chair: Julia Royall, Chief, International Programs, and Director, Malaria Research Telecommunications Network, for the National Library of Medicine, USA

This session will feature three panelists, all working with various tools and information technology to manage data to improve health in Africa.

Allen Hightower is Chief, Data Management Activity at CDC's National Center for Infectious Disease and a pioneer in initiating NLM's malaria research network at a remote site on Lake Victoria in Kenya. He has developed several tools for data collection and management which will change the speed and quality of data collection in Africa.

From the KEMRI-Wellcome Trust research unit on the coast of Kenya comes Tom Oluoch, systems operator/data manager and co-creator of a virtual library for this site, which brings together researchers from Kenya Medical Research Institute and Oxford University. His eyewitness case study is full of concrete examples of how IT and data management have brought expansion and change to this remote research unit.

Bob Mayes is Chief, Health Informatics Section, Zimbabwe CDC AIDS Program. CDC's program of technical assistance to Zimbabwe focuses on strengthening surveillance and laboratory measures, scaling up promising prevention and care strategies, supporting behavior change communication projects, data mining, semantic management of data for systematic review, and promoting technology transfer.

The presenters will discuss individual examples and case studies, as well as talk about how these tools can facilitate the discovery process.

1. Field Data Collection for the Malaria Research Network in Kenya
Allen Hightower, Centers for Disease Control, USA

Allen Hightower is Chief, Data Management Activity at CDC's National Center for Infectious Disease and a pioneer in initiating NLM's malaria research network at a remote site on Lake Victoria in Kenya. He has developed several tools for data collection and management which will change the speed and quality of data collection in Africa. He is currently evaluating field data collection using paperless GPS/data collection systems via Pocket PC-based personal data assistants in two projects:

(1) collecting census and GPS data for a wash-durable bednet study area and
(2) conducting a survey in a 15 village area on bednet usage for linkage with other health-related data.

2. Eye witness account: the role of IT and data management in expansion and change at a remote research unit in Kenya
Tom Oluoch, KEMRI-Wellcome Trust, Kenya

From the KEMRI-Wellcome Trust research unit on the coast of Kenya comes Tom Oluoch, systems operator/data manager and co-creator of a virtual library for this site, which brings together researchers from Kenya Medical Research Institute and Oxford University. His eyewitness case study is full of concrete examples of how IT and data management have brought expansion and change to this remote research unit.

3. CDC in Zimbabwe: strengthening regional surveillance and laboratory measures, supporting infrastructure development and promoting technology transfer
Robert Mayes, CDC AIDS Program, Zimbabwe

Bob Mayes is Chief, Health Informatics Section, Zimbabwe CDC AIDS Program. CDC's program of technical assistance to Zimbabwe focuses on strengthening surveillance and laboratory measures, scaling up promising prevention and care strategies, supporting behavior change communication projects, data mining, semantic management of data for systematic review, and promoting technology transfer.

4. Complex Data From Health Research
Themba Mohoto, Reproductive Research Unit, Chris Hani Baragwanath Hospital, Oweto

In the continuing search for better health for all, health researchers are faced with numerous methodological problems of a complex nature in their efforts to strengthen health programs, evaluate health systems and measure the impact of interventions. This in turn has posed greater challenges for data analysts.

This paper investigates the types of data produced in health research including:

1. Multi-stage survey data, e.g. Demographic Health Survey (DHS) in which data is collected at many levels such as household data, women data and children data and there is a need to link the data from these various levels.
2. Longitudinal or Repeated measures studies. Such data can arise either from cohort studies or from clinical trials. In this type of study there are repeated observations within individuals.
3. In clinical trial databases there are also difficulties with recording adverse events or concomitant medications, as there will be a variable number of these per patient.
4. A new area is that of cluster randomized trials which combines features of multistage sample data with features of clinical trial data.

Statisticians in this area are investigating ways of dealing with these problems.

Track III-D-6: Données & Santé : utilisations et enjeux (Data and Health: Usage and Issues)

Chair: Daniel Laurent, Université MLV, France ; Jean-Pierre Caliste UTC, France

Le poids du secteur de la Santé dans l'économie mondiale est devenu déterminant. Les dépenses pour la Santé représentent désormais 15% du PIB des Etats-Unis et 10% de celui de la France ou du Canada. Internet a bouleversé le domaine et accentué son ouverture au grand public en termes d'informations : il existe plus de 17 000 sites totalement dédiés à la santé et 40% des interrogations des internautes américains concernent des sites santé.

La diversité des situations médicales reposant sur l'utilisation de données complexes correspond a des angles d'approches variés : objectifs institutionnels et politiques, circulation d'informations médicales dans les réseaux spécialisés et par le biais d'internet, nouvelles pratiques médicales et assistance pour des sites isolés, évolution des services médicaux pour le praticien et le patient.

On constate une importance croissante des relations entre les composantes étatiques (systèmes de Sécurité sociale) et privées (assureurs, laboratoires pharmaceutiques…) tant au plan organisationnel (niveau de définition des actes) qu'à celui des aspects micro et macroéconomiques.

L'utilisation de données complexes (numériques, imagerie …) et la gestion des connaissances fait appel aux techniques de traitement de données de nature hétérogène en s'appuyant sur des approches résolument pluridisciplinaires venant conforter la théorie de l'information.

A partir de ces constatations, Codata France a fait du domaine de la santé l'un de ses trois axes prioritaires d'activité. Il propose d'organiser un atelier thématique sur cette question. Si nécessaire, il pourrait être réparti en 2 sessions spécialisées. Pourraient y être présentées les thématiques suivantes :

- les réseaux d'information ou les " autoroutes " de l'information en santé.
- données et internet (e health) : fiabilité, validité…
- les réseaux de santé et les réseaux coordonnés de soins : de nouveaux enjeux pour le " managed care "
- les systèmes d'information en santé : réseaux nationaux ou régionaux, réseaux d'établissements, réseaux de santé, cabinet médical…
- les enjeux de la télémédecine
- l'utilisation des données par centres d'appels (" call centers ")
- le dossier médical du patient
- la qualité des données
- la protection et l'archivage des données
- le confidentialité des données
- l'intéropérabilité des données et des systèmes

1. New information systems for the public healthcare insurance organization : the Catalan Health Service (CatSalut) in Spain

TORT I BARDOLET (Jaume), Generalitat de Catalunya, CatSalut, Barcelona, Spain

Key words : information systems, health care organisation, insurance, risk management, data.
Mots clés : systèmes d'information, management des systèmes de santé, gestion du risque, données.

Ten years after its creation, the Catalan Health Service (SCS) is initiating a reorientation process aimed at consolidating its role as the public healthcare insurance organization for all citizens of Catalonia. This reorientation involves generating a series of actions oriented more towards attention to the insurance holder/citizen, while maintaining a close relation with the suppliers of healthcare services from the public network.

This transformation coincides with the intention of generating qualitative and quantitative advancement regarding the structure of information systems available up to now. Thus a new Systems Plan is being drawn up, oriented towards the SCS's function as a public healthcare insurance organization.

1. Definition of the SCS's management needs

Aims:
1.1 To manage resources efficiently
1.2 To implant processes for continued improvement in service quality
1.3 To bring about active client management
1.4 To manage risk
1.5 To implant efficient administrative processes

These aims involve a series of needs that must be taken into account when developing new management and information systems.

- To back the management aims of the major working areas: demand, offer and internal administration, and lines of action for each of these (services).
- To facilitate the systematic drawing up of management reports based on parameters enabling the executive structure to make decisions concerning steps to be taken.
- To collect all necessary information properly and in good time by means of the most adequate software.

In order to specify these aims, a series of management levers has been devised:

To manage resources
To provide activity follow-up
To provide cost follow-up
To manage the quality level
To establish communication with clients
To manage risk
To improve the health of the population
To rationalize processes
To improve claim procedure for damages

Moreover, this has to be specified using pre-established follow-up parameters for drawing up the management reports.

2. Evaluating the developments and structure required

The proposal for the basic structure of the new information systems is based on three concepts and their corresponding identifiers:

- The insurance holder = personal identification code (CIP)
- The service providing unit = productive unit code (UP)
- The service / activity = service code

It has been planned that the different computer applications will work on a large data warehouse that will compile all activity (contracts of insurance holders with the productive structure) and which must make possible the generation of different views for each of the functions (see Graph 1).

The system has been graphically represented as a pyramid divided transversally into three parts. The lower trapezium shows the database (information) ; the middle, the computer applications (the treatment of information); and the upper triangle, the management information system.

The design of the information system is structured around four basic areas: demand, offer, activity and economy-finance (see Graph 2).

---

2. The planning and management of emergency treatment in Catalonia, by means of a specific information system

TORT I BARDOLET (Jaume), Generalitat de Catalunya, CatSalut, Barcelona, Spain

Key words : information systems, emergency, health care organisation, planning, data.
Mots clés : systèmes d'information, urgences, management des systèmes de santé, planification, données.

The Overall Emergency Plan has been used in Catalonia for the past three years. This is a global scheme that includes planning, precaution and prevention, management and supervision of emergency healthcare attention. It was created, above all, for those times of the year when there is an increased demand for healthcare attention for a variety of reasons.

The Plan includes:

- The analysis of the population requiring emergency attention: user characteristics, reasons for the examination, analysis of user expectations and motivations.
- Preventative actions: increased homecare coverage, increased influenza vaccine coverage, follow-up of users who have repeatedly requested emergency attention.
- Organizational actions: the drawing up by the hospitals of annual working plans for emergency attention, telephone-based back-up for mental health professionals, and coordination among healthcare mechanisms.
- An increase in the offer of contracted hospital discharges, and reinforcements in the summer and during periods of sustained growth in demand.

The information system
The Overall Emergency Plan is based on a specific information system -extranet- which makes it possible for a group of productive units from different healthcare areas to register - on a daily basis - emergency activity data from their centers, as well as other relevant information that allows the forecasting of increased demand and the quick and effective adoption of corrective measures. The extranet includes information regarding:

- Specialized attention (hospitals):
    - data concerning activity: emergency cases attended and admitted, hospital admissions, discharges and transfers to other centers
    - data concerning resource availability: patients awaiting admission, waiting period, available beds
- Continued primary attention: emergency activity of these centers
- Primary attention: data concerning continued attention, number of house calls
- Specific emergency services (061): number of telephone calls attended and services carried out.

---

3. Etude d'un système d'aide à la gestion de l'information dans la santé — Appliqué au domaine cardiovasculaire
Elisabeth Scarbonchi, Daniel Laurent, Christian Recchia, Université de Marne-la-Vallée, Institut Francilien d'Ingénierie des Services (I.F.I.S.), France

Dans le cadre d'un réseau de soins, le praticien et l'usager ont accès à un ensemble d'informations le concernant. Les informations sont réparties dans différents services d'un même hôpital voir plusieurs établissements. Dans ce système intervient la nature (typologie des informations), leur localisation et les volumes concernés, notamment les données informatiques lorsqu'il s'agit d'imagerie médicale.

Les réseaux à haut débit sont de nature à offrir des possibilités de connexion entre ces différentes sources pour une exploitation optimales dans les services de cardiologie.

Lorsqu'il s'agit de données numériques et textuelles, les techniques de datamining et textmining pourront être utilisés dans le but de produire de l'information à valeur ajoutée dans le cadre d'un fonctionnement opérationnel voir dans un contexte de recherche.

Lorsqu'il s'agit de sources d'images leur mise à disposition immédiate et interactive offre des possibilités et des perspectives d'animation et représentation dans un contexte opérationnel.

La mise en place d'un système d'informations multisources en réseau nécessitera de traiter avec une attention particulière les problèmes de sécurité et de propriété de données.

4. Données et santé : propriété, accès, protection, transmission. Les enjeux des réseaux de santé.
Christian Bourret, Université de Marne-la-Vallée, France
Serge Chambaud, Institut National de la Propriété Industrielle (I.N.P.I.), France
Elisabeth Scarbonchi, Université de Marne-la-Vallée, France
Daniel Laurent, Université de Marne-la-Vallée, France

Mots clés : propriété des données, confidentialité, dossier médical patient, réseaux de santé, autoroutes de l'information.
Key words : data ownership, confidentiality, medical record, business methods, patents, health care management, information networks.

La propriété et la protection des données constituent un des enjeux majeurs de la société post-industrielle fondée sur les biens immatériels : les services et la diffusion de l'information. Dans le contexte du développement de l'industrie de l'information, les données médicales constituent un enjeu commercial très important. Ces données sont très spécifiques. Il s'agit avant tout de données personnelles, sensibles et confidentielles, faisant l'objet de législations particulières. Pour bâtir notre problématique, nous nous appuierons sur l'expérience française des réseaux de santé, que nous élargirons ensuite à des comparaisons avec les Etats-Unis.

Le premier enjeu étudié sera celui de la propriété et de l'usage des données produites par les réseaux de santé. Nous l'analyserons à partir du dossier médical patient. Les données qu'il renferme appartiennent-elles au patient ? Aux différents médecins et aux organisations (hôpitaux, cliniques, assurance maladie …) pris individuellement ? Au réseau ? A l'entité qui l'héberge : notaire de l'information, infomédiaire ou hébergeur ? La réponse est loin d'être évidente. L'ensemble des données : le dossier global partagé, constitue-t-il en termes de propriété un tout différent de la somme de ses parties ? Peut-on vraiment strictement séparer l'usage des données de leur propriété ? Nous analyserons les différentes réponses actuelles possibles à ces questions.

Nous évoquerons ensuite une autre question déterminante : l'accès du patient à la consultation de ses données de santé personnelles. En France, la nouvelle loi du 4 mars 2002 a posé de grands principes mais a laissé de nombreuses interrogations en suspens. Cet accès se fera-t-il directement ? Indirectement par le biais d'un médecin ? Et à quelles données le patient aura-t- il accès ? A l'intégralité de son dossier ou à un résumé ? Aura-t-il également accès aux commentaires des praticiens ? Nous tracerons des pistes de réflexion pour éclairer toutes ces questions.

Tout se complique encore quand, comme c'est largement le cas aux Etats-Unis, les patients constituent leur propre dossier médical. Dans ce cas, quelle en est la fiabilité ? Peut-il être utilisé par des professionnels qui engageraient ainsi leur responsabilité ?

En terme de propriété industrielle et intellectuelle, se pose aussi la question de la brevetabilité et de la protection des logiciels de création, de gestion ou de diffusion du dossier médical patient. Les dossiers médicaux patients sont-ils protégeables ? Les critères de brevetabilité classiques s'appliquent-ils ou non à eux ? Ou bien, constituent-ils des " business methods " et, dans ce cas, comment les protéger ? Les réponses peuvent varier selon les pays. Nous aborderons ces questions à travers une comparaison entre les possibilités offertes en France et aux Etats-Unis.

La transmission des données médicales constitue un autre enjeu majeur, celui des autoroutes de l'information. Nous examinerons deux aspects essentiels de l'évolution actuelle, notamment en France : l'effacement progressif de l'Etat au profit d'acteurs privés et le choix fondamental entre la sécurisation d'un réseau de transmission de données médicales (Réseau Santé Social de Cégétel-Vivendi) ou de la sécurisation des données elles-mêmes (France Télécom ou Cegedim).

5. Les réseaux de santé : une expérimentation française centrée sur le partage de l'information
Gabriella Salzano, Université de Marne-la-Vallée, France
Christian Bourret, Université de Marne-la-Vallée, France
Jean-Pierre Caliste, Université de Technologie de Compiègne (UTC), France
Daniel Laurent, Université de Marne-la-Vallée, France

Mots clés : réseaux de santé, systèmes d'information, information partagée.
Key words : health care management, information systems, data, shared information.

Depuis le début des années 1980, l'ensemble des grands pays industrialisés sont confrontés au problème de la maîtrise des coûts de leurs systèmes de santé et en particulier de ceux de l'hospitalisation. Une solution envisagée a été le " virage ambulatoire " visant à privilégier la médecine de ville en s'appuyant sur les nouvelles technologies de l'information et de la communication (NTIC). En France, une voie originale a été expérimentée : les réseaux de santé. Elle a été légitimée par la loi du 4 mars 2002 relative au droit des malades et à la qualité du système de santé.

Les réseaux de santé se veulent résolument au service du patient. Leurs objectifs sont de décloisonner le système de santé en améliorant l'indispensable relation ville-hôpital mais aussi les relations entre les différents professionnels en charge du même patient. Il s'agit d'assurer la qualité et la continuité de soins par la mise en place d'une organisation innovante, fondée sur des valeurs partagées, comme la construction de pratiques collégiales et non plus individuelles ou hiérarchisées, et un meilleur partage de l'information.

Les systèmes d'information constituent le pivot des réseaux de santé. Ils doivent tout d'abord assurer l'interopérabilité (coordination et intégration) de différents autres sous-systèmes, notamment les systèmes d'information propres aux hôpitaux ou cliniques, les logiciels de gestion de cabinets médicaux ou des autres professionnels. Il doivent aussi permettre l'accès à des bases de données ou à des logiciels d'aides à la décision (référentiels …) comme aux services de télémédecine. Ils doivent aussi assurer la gestion de services spécifiques au réseau : plate-forme d'orientation des urgences et / ou centre d'appels, dossier patient partagé au sein du réseau … Nous analyserons les principaux problèmes à résoudre, en termes d'organisation et d'applications.

Les réseaux de santé répondent à des forts besoins de changement. Leur mise en place et leurs performances doivent être évaluées. L'évaluation influence fortement l'élaboration du système d'information, car celui-ci devra fournir les données indispensables au suivi des indicateurs d'évaluation et répondre à des exigences de qualité, spécifiques aux objectifs des réseaux.

Dans cette communication, nous évoquerons les enjeux et les méthodologies d'évaluation des réseaux de santé, en soulignant les interactions avec les méthodologies d'élaboration des systèmes d'information, dans un cadre de management de projets complexes.

Behavioral and Social Science Data

Track I-C-4: Government as a Driver in Database Development in the Behavioral Sciences

Chair: David Johnson, Building Engineering and Science Talent, USA

The behavioral sciences have not had a tradition of data sharing. Thus they have been somewhat behind other sciences in the development of databases. Officials in several science agencies of the US federal government have been concerned about this lack of data sharing and have taken measures to stimulate development. The purpose of this panel is to explore the ways that government agencies can arrange funding opportunities to stimulate innovation in areas that scientists within given fields have been reluctant to address.The work of three US agencies will be highlighted: The National Science Foundation, the National Institutes of Health, and the Federal Aviation Administration.

Government and science often exercise reciprocal influences on each other. The three examples that that will be explored in this panel session represent three discrete models by which governments may stimulate a science to produce knowledge in a way that it would not have in the absence of the government's effort.

---

1. Sharing data collection and sharing collected data: The NICHD Study of Early Child Care and Youth Development
Sarah L. Friedman, The NICHD Study of Early Child Care and Youth Development, USA

The NICHD Study of Early Child Care and Youth Development came to life as a result of a 1988 NICHD solicitation (RFA) and is scheduled to terminate at the end of 2009. The aim of the solicitation was to bring together investigators from different universities or research institutions to collaborate with NICHD staff on the planning and execution of one longitudinal study with data to be collected across sites. The idea for such a collaborative study was unprecedented in the scientific field of developmental psychology.

Ten data collection sites were selected on a competitive basis and the affiliated investigators, in collaboration with NICHD staff, have designed the different phases of the solicited longitudinal study and have implemented it. While the data collected at each of the sites belongs to the site, NICHD required that each of the 10 sites would send its data to a central location, the Data Acquisition and Analysis Center, for data editing, data reduction and data analyses. The study investigators, in collaboration of the data center staff, guide the data acquisition and analyses. Upon completion of an agreed upon quota of network authored scientific papers for a given phase of the study, individual study investigators get access to the data sets of the entire sample. A few months after the data sets and supporting documentation are available to individual study investigators for their exclusive use, the same data sets are made available to interested and qualified others in the scientific community.

While the archiving of the data is done by an NICHD grantee, the Murray Center at Radcliff College has expressed interest in archiving the data and supporting their use by interested and qualified investigators. If the grantee institutions will accept the Murray Center request, the data collected by the grantees will be available to the scientific community beyond the life of the grant.

2. Data Sharing at NIH and NIA
Miriam F. Kelty, National Institute on Aging, Office of Extramural Activities, USA

NIH published its policy mandating sharing of unique biological resources in 1986. Sixteen years later NIH published a draft policy. It states that NIH expects the timely release and sharing of final research data for use by other researchers. Further, NIH will require extramural and intramural investigators to promulgate a data sharing plan in their research proposals or to explain why a plan to share data is not possible. The policy is available for comment until June 1. The presentation will provide background information and summarize public comments.

NIA staff have been leading advocates for data sharing and have encouraged it among grantees, particularly when research involves large data sets that are valuable research resources and impractical to replicate. NIA will provide funds to make data that are well documented and user-friendly available to other researchers. Some examples of NIA supported activities in support of data sharing are described below:

The National Archive of Computerized Data on Aging (NACDA), located within the Interuniversity Consortium for Political and Social Research (ICPSR), is funded by the National Institute on Aging. NACDA's mission is to advance research on aging by helping researchers to profit from the under-exploited potential of a broad range of datasets. NACDA acquires and preserves data relevant to gerontological research, processing as needed to promote effective research use, disseminates them to researchers, and facilitates their use. By preserving and making available the largest library of electronic data on aging in the United States, NACDA offers opportunities for secondary analysis on major issues of scientific and policy relevance.

NACDA supports a data analysis system that allows the user to access subset variables or cases. The system can be used with a variety of data stets, including the Longitudinal Survey on Aging, National Survey of Self-Care and Aging, National Health and Nutrition Survey, National Hospital Discharge Survey, and the National Health Interview Survey.

NIA supports a range of studies that have agreed to make data available to researchers. An example is the Health and Retirement Study, a nationally representative study that collects data on aging and retirement. The study is based at the University of Michigan and the Michigan Center on Demography of Aging makes data available to a range of researchers. Some data is available to anyone for analysis while other data sets are restricted and require contractual agreements prior to being made available for use.

The presentation will address NIA's experience with the use of available data sets and raise some issues surrounding data sharing.

3. Data Archiving for Animal Cognition Research: The NIMH Experience
Howard S. Kurtzman, Cognitive Science Program, National Institute of Mental Health, USA

In July 2001, the National Institute of Mental Health (a component of the U.S. National Institutes of Health) sponsored a workshop on "Data Archiving for Animal Cognition Research." Participants included leading scientists as well as experts in archiving, publishing, policy, and law. Due to the focus on non-human research, participants were able to devote primary attention to important issues aside from protection of confidentiality, which has dominated most previous discussions of behavioral science archiving. The further limitation of the workshop's scope to animal cognition research allowed archiving to be examined realistically in the context of one particular scientific community's goals, methods, organization, and traditions.

The workshop produced a set of conclusions, detailed in a formal report, concerning: (1) the likely impacts of archiving on research and education, (2) guidelines for incorporating archiving into research practice, (3) contents of archives, (4) technical standards, and (5) organizational and policy issues. The presentation will review these conclusions and describe activities following up on the workshop. Also discussed will be the applicability of the workshop's conclusions to other areas of behavioral science and how this workshop's approach to stimulating archive development might serve as a model for other fields.

4. Data Sharing and the Social and Behavioral Sciences at the National Science Foundation
Philip Rubin, Division of Behavioral and Cognitive Sciences, USA

At the heart of the National Science Foundation's (NSF) strategic plan are people, ideas, and tools. In the latter area, our goal is to provide broadly accessible, state-of-the-art information-bases and shared research and education tools. We actively encourage data sharing across all of our fields of study. This presentation will provide examples from the social and behavioral sciences. As data sharing is encouraged and increased, however, there are growing concerns and issues related to privacy and confidentiality. These issues will also be discussed, as will future directions in information sharing.

At the NSF, the Directorate for Social, Behavioral, and Economic Sciences (SBE) participates in special initiatives and competitions on a number of topics, including infrastructure to improve data resources, data archives, collaboratories, and centers.

The breadth of fields is wide in our Directorate, ranging from Anthropology through Political Science and Economics. However, common to many of the disciplinary areas that we support is a rapid change in how the science is being done. What is emerging is a large scale social science, driven by computational progress, the need for scientific expertise across a number of domains, growing bodies of data and other information, and theoretical and practical issues that require for their understanding a broader view than has been taken in the past.

This change will be illustrated by some examples of recent or continuing projects that we are supporting. For example, physical anthropologists utilize tools from a wide range of overlapping disciplines ranging from molecular biology (population genetics) to field ecology to remote sensing (paleoanthropology). In all of these areas large amounts of data are generated that are conducive to the establishment of digital libraries, databases, web-based archives and the like. A recent SBE Infrastructure award will be described that supports a number of interrelated activities that will advance research in physical anthropology, evolutionary biology, neuroscience and any others that may require information and/or biomaterials from nonhuman primates.

An example in geography is the National Historical Geographic Information System (NHGIS) at the University of Minnesota, Twin Cities. This project upgrades and enhances U.S. Census databases from 1790 to the present, including the digitization of all census geography so that place-specific information can be readily used in geographic information systems. We expect that the NHGIS will become a resource that can be used widely for social science training, by the media, for policy research at the state and local levels, by the private sector, and in secondary education.

Last year the National Science Board approved renewal of NSF support for the Panel Study of Income Dynamics (PSID). The PSID is a longitudinal survey initiated in 1968 of a nationally representative sample for U.S. individuals and the family units in which they reside. The major objective of the panel is to provide shared-use databases, research platforms and educational tools on cyclical, intergenerational and life-course measures of economic and social behavior. With thirty-plus years of data on the same families, the PSID can justly be considered a cornerstone of the infrastructure support for empirically based social science research.

Additional examples abound, and will be discussed. These include CSISS, the Center for Spatially Integrated Social Science, at the University of Santa Barbara; the fMRI Data Center at Dartmouth College, a national cognitive neuroscience resource; data-rich linguistics projects that support both the preservation of knowledge of disappearing languages and statistically-guided approaches to increasing our understanding of ongoing language use; systems for storage and dissemination of multimodal (audio, visual, haptic, etc.) data; and systems and techniques for the meta-analysis of large scale data sets.

Data sharing is at the heart of NSF's mission and of our vision of the social and behavioral sciences. This presentation is intended to provide an overview of that vision.

Track I-D-6: Database Innovation in the Behavioral Sciences and the Debate Over What Should Be Stored

Session organizer: US National Committee for the International Union of Psychological Sciences, National Academy of Sciences, Washington, D.C., USA

Chair: Merry Bullock, American Psychological Association

Data sharing is not the norm in behavioral science, although there are pockets of change and innovation. At the same time, a debate is underway regarding what data from experiments are worth placing in databases to be available for others. As it becomes possible to store huge quantities of data, it is becoming more necessary to assure that databases grow into useful tools rather than clogged informational arteries. This panel has two objectives: to inform attendees of innovations and to discuss the possible criteria for determining what should be included in databases.

Panelists will discuss several innovative databases that are proving transformational for the fields they touch. For example, a database of functional magnetic resonance images of the brain created at Dartmouth College is making it possible to test hypotheses about brain-behavior relations on data pooled across many individual studies; a database of geographic information based at the University of California, Santa Barbara is allowing those in a variety of disciplines to look at the influence of location on such things as health behaviors, social development, and wealth accumulation. A database of aptitude test scores at the University of Virginia is a test bed for statistical innovations that are making it possible to legitimately compare data and not just outcomes from disparate studies.

The Panel will describe several of these innovations in behavioral and other sciences, and will address important emerging issues. For example, the fMRI database (originally envisioned as capturing all the images from most of the major neuroscience journals) is constrained because of file size-images from a single journal consume terabytes of storage space and raise important questions of accessibility. As the behavioral sciences evolve toward more common acceptance of data sharing, those in the behavioral sciences must evolve toward a more common understanding of what should be contained in a database and what sorts of data are appropriate for archiving. Examples and issues from other disciplines will help inform the discussion.

1. Acquisition Criteria at the Murray Research Center: A Center for the Study of Lives
Jacquelyn B. James, Murray Research Center

The Murray Research Center is a repository for social and behavioral sciences data on the in-depth study of lives over time, and issues of special concern to American women. The center acquires data sets that are amenable to secondary analysis, replication, or longitudinal follow-up. In determining whether or not to acquire a new data set for the archive, several kinds of criteria are used. The criteria can be roughly grouped into five general categories: content of the study, methodology, previous analysis and publication, historical value, and cost of acquiring and processing the data. Each of these will be described with an indication of the relative importance of each criterion, where possible.

2. What Functional Neuroimaging Data is 'Worth' Sharing and the Scope of Large-Scale Study Data Archiving
John Darrell Van Horn, The fMRI Data Center, Dartmouth College, USA

Functional neuroimaging studies routinely produce large sets of raw data that comprise both functional image time series as well as high-resolution anatomical brain volumes. It is often the case that these data are then passed through several steps of processing and then only a limited set of the statistical output is presented in papers published in the peer-reviewed literature. Arguments for archiving only these summary results have suggested that they are of greater value than that of the raw data itself. However, since with each step of processing the information content of a data set remains constant or is reduced, it is difficult to see the source of any increased scientific value. The fMRI Data Center (fMRIDC) strives to archive complete raw functional neuroimaging data sets accompanied by enough information that anyone else would be able to reconstruct the steps in processing of the data and arrive at the same statistical brain map as the original authors. To achieve this, the fMRIDC requests that authors of published studies provide considerably more study 'meta' and raw data than is typically presented in their published article. As such, several studies currently in the fMRIDC archive rival the size of the entire human genome database (~20GB compressed). Through the storing of complete study data sets, the fMRIDC effort will serve to not only advance thinking into fundamental concepts about brain function by permitting others to examine the published neuroimaging data of others but also to document more thoroughly the scientific record of work in the fields of functional brain imaging and cognitive neuroscience.

3. Accession and Sharing of Geographic Information
Michael F. Goodchild, University of California, Santa Barbara, USA

Geographic information is a well-defined type, with complex uses and production systems. The Alexandria Digital Library began as an effort to provide remote access to a large collection of geographic information (maps and images), but has evolved into a functional geolibrary (a digital library that can be searched using geographic location as the primary key). I use ADL to illustrate many of the issues and principles inherent in sharing geographic information, and in policies regarding its acquisition by archives, including granularity, metadata schema, support for search across distributed archives, portals and clearinghouses, and interoperability.

Informatics and Technology

Track I-C-5: Data Archiving

Chair: Seamus Ross

---

1. Report of Activities of the CODATA Working Group on Archiving Scientific Data
William Anderson, Praxis101, Rye, NY, USA
Steve Rossouw, South African National Committee for CODATA
Co-organizers: CODATA Working Group on Archiving Scientifc Data

A Working Group on Scientific Data Archiving was formed following the 2000 International CODATA Conference in Baveno, Italy. The Working Group has (1) built a list of anotated primary references to published reports and existing scientific data archives, (2) constructed a classification scheme to help organize and expose the many issues and requirements of archiving, preserving, and maintaining access to scientific and technical data, (3) helped sponsor a workshop in South Africa on archiving scientific and technical data, and (4) proposed collaborating with the International Council for Scientific and Technical Information (ICSTI) to build and maintain an internet portal focused on scientific data and information archiving, preservation and access. The objectives of these efforts is to provide scientists and scientific data managers a framework of information and references that can assist in securing the resources and commitments needed to preserve and archive scientific data. This presentation outlines the results of these efforts with the goal of stimulating discussion of the organizing framework as well as the definitions and relationships among identified issues.

---

2. The NIST Data Gateway: Providing Easy Access to NIST Data Resources
Dorothy M. Blakeslee, Angela Y. Lee, and Alec J. Belsky, National Institute of Standards and Technology, USA

The National Institute of Standards and Technology (NIST) maintains a wide range of scientific and technical data resources, including free online data systems and PC databases available for purchase. However, many people are not familiar with these various NIST data collections and the types of data they contain. To help scientists, engineers, and the general public find out quickly and easily whether data they need are available at NIST, NIST has built a web portal to NIST data resources. The first version of this portal, the NIST Data Gateway (http://srdata.nist.gov/gateway), provides easy access to 26 online NIST data systems and information on 48 NIST PC databases. NIST Data Gateway users can specify a keyword, property, or substance name to find the NIST data resources that contain standard reference data meeting their search criteria. When users find a data resource they want to use, links are provided so they can access or order that resource. In this paper, we describe how version 1.0 of the NIST Data Gateway was built and discuss some of the issues that arose during the design and implementation stages. We include experience we gained that we hope will be useful to others building data portals. We also discuss future plans for the NIST Data Gateway, including efforts to provide access to additional NIST data resources.

3. Long Term Data Storage: Are We Getting Closer to a Solution?
A. Stander and N. Van der Merwe, Department of Information Systems, University of Cape Town, South Africa
Steve F. Rossouw, South Africa National Committee for CODATA, South Africa

Many scientific and socioeconomic reasons exist for the long term retention of scientific and lately also business data. To do so successfully, the solution must be affordable and also technologically flexible enough to survive the many technology changes during its useful life. This paper looks at the current status of available technology for long term data storage, more specific the standards that exist for data interchange, the creation and storage of metadata, data conversion problems and the reliability and suitability of digital storage media. Even if in the ideal format, application and database management software is needed to store and retrieve the data.

Typically the life expectancy of such software is much shorter than that of the storage media and as this has already been the cause of major data loss, possible solutions are investigated. Most research into long term data storage focus on large to very large databases. It is often forgotten that small, but very important pockets of scientific data exist on the computers of individual researchers or smaller institutions. As most of the time this is stored in application specific formats with a short lifespan, strategies for the preservation of smaller amounts of data are also looked at.

4. Prototype of TRC Integrated Information System for Physicochemical Properties of Organic Compounds: Evaluated Data, Models, and Knowledge
Xinjian Yan, Thermodynamics Research Center (TRC), National Institute of Standards and Technology, USA
Qian Dong, Xiangrong Hong, Robert D. Chirico and Michael Frenkel

Physicochemical property data are crucial for industrial process development and scientific research. However, such data that have been experimentally determined are not only very limited, but also deficient in critical evaluations. Moreover, the models developed for the prediction of physicochemical property have rarely been presented with sufficient examination. This situation makes it very difficult to understand the data that are obtained from reference books, databases or models after a time-consuming effort. Therefore, we aim at developing a comprehensive system, TRC Integrated Information System (TIIS), which consists of evaluated data, models, knowledge and functions to infer, and then to recommend, the best data and models. Additionally, it provides valuable information for users to have a better understanding of physicochemical property data, models, and theory.

Evaluated physicochemical property data in TIIS are mainly selected from the TRC Source data system, which is an extensive repository system of experimental physicochemical properties and relevant measurement information. Data uncertainty and reliability are analyzed based on scientific data principles, statistics, and highly evaluated property models. Information about experimental condition, data processing, etc., is recorded in a detailed way.

Reliability of the data predicted by a model cannot be determined without a full description of the model's ability.

Each model in TIIS is carefully examined by using evaluated data, with emphasis on the predictive ability for calculating the compounds not used in processing the model's parameters, and applicable compound classes, for which the model can produce reasonably good property data. For a given compound, the best predictive value is recommended according to models' performances in calculating evaluated data set. TIIS also provides regression analyses and optimization functions so that users are able to process model parameters by using the current best experimental data set for a particular compound.

A property value, a model or a chemical system cannot be fully understood without sufficient supporting information. Therefore, the knowledge that describes characteristics of property data, models, molecular structures, and the results from theoretical analysis and calculation, is provided by TIIS.

5. An Introduction of CODATA-China Physical and Chemical Database Information System

Yan Baoping, Director, Computer Network Information Center, CAS, China
Xiao Yun, Secretary General, Chinese National Committee for CODATA, China
Zhang Hui, Secretary, Chinese National Committee for CODATA, China
Jin Huanian, Engineer, Computer Network Information Center, CAS, China

In 2001 the Chinese Ministry of Science and Technology made the decision to bring the data center coordinated by CODATA-China into the basic work of the National Key Research Development Program, rendering long-term support for the accumulation, development and utilization of the technological basic data work by starting the special technological basic project.

A database information service system is expected to be set up within 3 to 5 years with the CODATA-Chian Physical and Chemical Database Information System as the main body, involving the subjects of agriculture, forestry, mechanism, material, biology, etc., so as to form a centered group of CODATA-China Physical and Chemical Database Information System, which, targeting the field of mathematics, physics, and chemistry, is able to provide basic and applied data for the scientific research and production. At present the data contained in CODATA-China Physical and Chemical Database Information System mainly includes: the Chinese nuclear data, the Chinese atom and molecule data, the Chinese chemistry and chemical industry data, the geothermodynamics data, the chemdynamics data, the Chinese aviation material data, and the Chinese feedstuff technology data. The Computer Network Information Center, CAS, will work as the general center,providing this project with service platform and technologic support based on centralized management assisted by distributed management. Relying on high-performance Unix server and the database management system of Oracle, the data application service platform of superb usability and efficiency will be developed based on the high-performance and transplantable software development language of JAVA. The advanced full text retrieval system in China, the TRS Full Text Retrieval System, will be used to provide highly efficient and reliable service of full text data retrieval, and the data service will be realized in the Web mode through Internet.

Track I-C-6:
Ingénierie de la veille technologique et de l'intelligence économique (Data for Competitive Technical and Economic Intelligence)

Chair: Clément Paoli, Université MLV, France

La production d'information élaborée à partir de l'analyse mathématique et linguistique des sources d'information électroniques contenant des données scientifiques factuelles et technologiques textuelles, constitue la matière première des décisions stratégiques.

Le développement des méthodes et outils logiciels permettant un criblage systématique des sources d'information provenant des banques de données en ligne et d'Internet permet d'obtenir des corpus d'information à forte valeur ajoutée.

Le management de la connaissance repose sur l'obtention rapide et de qualité de données et d'information élaborée. Les techniques d'amélioration de ces données sont souvent associées avec les logiciels de traitement retenus pour les problèmes d'intelligence économiques.

Les problèmes de standardisation et d'interopérabilité des systèmes locaux entre eux et avec l'information externe constituent des bases de discussions et d'échanges de vue rechercher dans cette session.

Les principaux thème proposés ici sont exposés pour le ouvrir le débat, d'autres propositions seront évaluées.

- Accès aux sources d'information : moteurs d'interrogation
- Représentation des connaissances : traitements sémantiques - cartographie - imagerie
- Méthodes et outils d'analyse statistique : analyse statique et en ligne
- Analyse mathématiques de données : AFC, Classifications hiérarchisées
- Méthodes et outils linguistiques : extraction terminologique (analyse sémantique)
- Data Mining , Texte Mining : interopérabilité et bases hétérogènes clustérisation
- Knowledge Management : l'information comme ressource
- Vulnérabilités informationnelles : qualité des données et des informations, protection

---

1. Ingénierie de la veille pédagogique et gestion des connaissances en enseignement supérieur (Data for competitive pedagogy and knowledge management in higher education)
Jean-Paul Pinte, Université Marne La Vallée, France

L'enseignement universitaire est condamné à se renouveler, à redéfinir ses paradigmes, sinon il se sclérose.

Des changements pour ces derniers sont apparus depuis quelques années dans de nombreux domaines comme le téléphone sans fil, la médecine préventive, l'écologie, la mondialisation…

Pour ce qui est de l'enseignement nous sommes entrés dans le paradigme de l'apprentissage.
Les pressions nous viennent principalement du monde du travail avec entre autres la création de nouveaux environnements de travail, l'apparition de nouvelles caractéristiques de clientèles, une explosion des connaissances et des ressources, le développement fulgurant des Technologies de l'Information et de la Communication, et, surtout, l'arrivée de nouveaux étudiants de tout âge, de toute provenance, avec des motivations et des compétences diversifiées à l'extrême.

En dehors de l'enseignement de matières au niveau le plus élevé de la connaissance, de la recherche et de la production de savoirs, l'université assure aujourd'hui un troisième rôle économique et social ayant pour objectif la production de valeur ajoutée et débouchant sur la recherche finalisée.

L'économie du savoir supplante l'économie matérielle et les universités sont de plus en plus "entrepreneuriales".

Une remise en cause profonde de l'université est déjà en cours. Elle vise à s'ouvrir à ce nouveau rôle par la mise en place de pédagogies " actives ", de formations ouvertes et à distance (E-learning, campus virtuels, numériques, …).

Avec les TIC on ne peut plus enseigner comme avant.

Des TIC aux TIC, il nous faut maintenant passer à " Technologies pour l'Intelligence et la Connaissance " .

La veille pédagogique est une des principales clés de réussite pour accompagner ce changement.

---

2. L'analyse des mots associés pour l'information non scientifique
(Co-Word analysis for non scientific information)
Bertrand Delecroix and Renaud Eppstein, ISIS/CESD, Université de Marne La Vallée, France

Co-word analysis is based on a sociological theory developed by the CSI and the SERPIA (Callon,Courtial,Turner,Michelet) in the middle of the eighties. It measures association strength between terms in documents to reveal and visualise evolution of science through the construction of clusters and strategic diagram. Since, this method has been successfully pplied to investigate the structure of many scientific fields. Nowadays it occurs in many software systems which are used by companies to improve their business and define their strategy but the relevance in this kind of application has not been proved yet.

Through the example of economic and marketing information on DSL technologies from Reuters Business Briefing, this presentation gives an interpretation of co-word analysis for this kind of information. After an outlook of the software we used (Sampler and LexiMine) and after a survey of the experimental protocol, we investigate and explain each step of the co-word analysis process : terminological extraction, computation of clusters and strategic diagram. In particular, we explain the meaning of every parameter of the method : the choice of variables and similarity measures is discussed. Finally we try to give global interpretation of the method in an economic context. Further studies will be added to this work in order to allow a generalisation of these results.
Keywords : clustering, co-word analysis, competitive intelligence

---

3. Stratégies du partenariat scientifique entre les pays de l' UE et les pays en développement : indicateurs bibliométriques
P.L. Rossi , IRD, Centre d'Ile de France, Bondy, France

L'exploitation de la base de données bibliographique Science Citation Index (SCI) de l'ISI (Institute for Scientific Information, Philadelphie) permet de concevoir des indicateurs bibliométriques servant à caractériser les stratégies de partenariat scientifique qui existent entre les pays de l'Union européenne et les pays en développement.

Les données disponibles dans la base SCI que nous avons exploité permettent d'établir de multiples indicateurs concernant les productions scientifiques des pays, de leurs régions, de leurs institutions, des politiques scientifiques nationales, des proximités et des affinités entre différents acteurs. Ils sont regroupés dans des indicateurs de productions scientifiques, des indicateurs de spécialisation, des indicateurs relationnels.

Cette étude a été réalisée sur les données bibliographiques de la période 1987-2001 et concerne les pays de l'Union européenne en plus des pays de l'Afrique, de l'Amérique latine ainsi que de l'Asie.

En ce qui concerne les stratégies de partenariat scientifique entre les 15 pays de l'Union européenne et les pays du continent africain, trois " grandes " catégories de partenaires européen peuvent être définies :

- les pays de l'Union européenne qui ont un profil de partenariat " proche " des principaux producteurs scientifiques africains : l'Autriche, l'Allemagne, l'Espagne, la Finlande et l'Italie,
- les pays de l'Union européenne qui ont un profil de partenariat avec un " engagement " fort en Afrique subsaharienne : le Danemark, les Pays Bas et la Suède,
- les pays de l'Union européenne qui ont un profil de partenariat avec des pays pour lesquels des relations liées à l'histoire coloniale et à la langue existent : la Belgique, la France et la Grande Bretagne.

Pour ces deux derniers pays est à signaler la différence de l'impact qu'ils ont sur les productions nationales de leurs principaux partenaires : très important pour la France, plus modéré pour la Grande Bretagne.

---

4. La fusion analytique Data/Texte : nouvel enjeu de l'analyse avancée de l'information (The merging of structured and unstructured data : the new challenge of advanced information analytics)
J.F. Marcotorchino, Kalima Group

Track III-C-4: Attaining Data Interoperabilty

Chair: Richard Chinman, University Corporation for Atmospheric Research, Boulder, CO, USA

Interoperability can be characterized as the ability of two or more autonomous, heterogeneous, distributed digital entities (e.g., systems, applications, procedures, directories, inventories, data sets, ...) to communicate and cooperate among themselves despite differences in language, context, or content. These entities should be able to interact with one another in meaningful ways without special effort by the user - the data producer or consumer - be it human or machine.

By becoming interoperable the scientific and technical data communities gain the ability to better utilize their own data internally and become more visible, accessible, usable, and responsive to their increasingly sophisticated user community. When "the network is the computer", interoperability is critical to fully take advantage of data collections and repositories.

Two sets of issues affect the extent to which digital entities efficiently and conveniently interoperate: syntactic and semantic interoperability.

Syntactic interoperability involves the use of communication, transport, storage and representation standards. For digital entities to interoperate syntactically, information (metadata) about the data types and structures at the computer level, the syntax of the data, is exchanged. However, if the entities are to do something meaningful with the data, syntactic interoperability is not enough, semantic interoperability is also required.

Semantic interoperability requires that a consistent interpretation of term Usage and meaning occur. For digital entities to interoperate semantically, consistent information (metadata) about the content of the data - what the basic variable names are and mean, what their units are, what their ranges are - is exchanged. This information can be referred to as the semantic search metadata, since it can be used to search for (and locate) data of interest to the user. However, this is not the metadata that is required for semantic interoperability at the data level, although it does contain some of the same elements. The semantic information required for machine-to-machine interoperability at the data level is the information required to make use of the data. For example, the variable T is sea surface temperature, the data values correspond to ºC divided by 0.125, missing values are represented by -999, ... This information can be referred to as the semantic use metadata. Without this information, the digital entity, be it machine or application, cannot properly label the axes of plots of the data or merge them with data from other sources without intervention from a knowledgeable human.

There are other sets of issues that affect interoperability:

- Political/Human Interoperability
- Inter-disciplinary Interoperability
- Legal Interoperability
- International Interoperability

This session is about all sets of interoperability issues, but especially focused on attaining semantic interoperability at the data level.

1. Interoperability in a Distributed, Heterogeneous Data Environment: The OPeNDAP Example
Peter Cornillon, Graduate School of Oceanography, University of Rhode Island, USA

Data system interoperability in a distributed, heterogeneous environment requires a consistent description of both the syntax and semantics of the accessible datasets. The syntax describes elements of the dataset related to its structure or organization, the contained data types and operations that are permitted on the data by data system elements. The semantics give meaning to the data values in the dataset. The syntactic and semantic description of a dataset form a subset of the metadata used to describe it; other metadata often associated with a dataset are fields that describe how the data were collected, calibrated, who collected them, etc. Although important, indeed often essential, to meaningfully interpret the data, these additional fields are not required for machine-to-machine interoperability (Level 3 Interoperability) in a data system. We refer to semantic metadata required to locate a data source of interest as semantic search metadata and semantic metadata required to use the data, for example to label the axes of a plot of the data or to exclude missing values from subsequent analysis, as semantic use metadata.

In this presentation, we summarize the basic metadata objects required to achieve Level 3 Interoperability in the context of an infrastructure that has been developed by the Open source Project for a Network Data Access Protocol (OPeNDAP) and how this infrastructure is being used by the oceanographic community in the community-based National Virtual Ocean Data System (NVODS). At present, in excess of 400 data sets are being served from approximately 40 sites in the US, Great Britain, France, Korea and Australia. These data are stored in a variety of formats ranging from user developed flat files to SQL RDBMS to sophisticated formats with well defined APIs such as netCDF and HDF. A number of application packages (Matlab, IDL, VisAD, ODV, Ferret, ncBrowse and GrADS) have also been OPeNDAP-enabled allowing users of these packages to access subsets of data sets of interest directly over the network.

2. Interoperable data delivery in solar-terrestrial applications: adopting and evolving OpENDAP
Peter Fox, Jose Garcia, Patrick West, National Center for Atmospheric Research, USA

The High Altitude Observatory (HAO) division of NCAR investigates the sun and the earth's space environment, focusing on the physical processes that govern the sun, the interplanetary environment, and the earth's upper atmosphere.

We present details on how interoperability within a set of data systems support by HAO and collaborators has driven the implementation of services around the Data Access Protocol (DAP) originating in the Distributed Oceanographic Data System (DODS) project. The outgrowth of this is the OpENDAP - an open source project to provide reference implementations of the DAP and its core services.

We will present the recent design and development details of the services built around the DAP, including interfaces to common application programs, like the Interactive Data Language, the web, and server side data format translation and related services.

We also present examples of this interoperability in a number of science discipline and technology areas: the Coupling, Energetics and Dynamics of Atmospheric Regions (CEDAR) program, the Radiative Inputs from Sun to Earth (RISE) program, the Earth System Grid II project, and the Space Physics and Aeronomy Collaboratory.

3. The Earth System Grid: Turning Climate Datasets Into Community Resources
Don Middleton, NCAR, Boulder, CO, USA
Ethan Alpert, NCAR, Boulder, CO, USA
David Bernholdt, Oak Ridge National Laboratory, Oak Ridge, TN, USA
David Brown, NCAR, Boulder, CO, USA
Kasidit Chancio, Oak Ridge National Laboratory, Oak Ridge, TN, USA
Ann Chervenak, USC/ISI, Marina del Ray, CA, USA
Luca Cinquini, NCAR, Boulder, CO, USA
Bob Drach, Lawrence Livermore National Laboratory, Livermore, CA, USA
Ian Foster, Argonne National Laboratory, Argonne, IL, USA
Peter Fox, NCAR, Boulder, CO, USA
Jose Garcia, NCAR, Boulder, CO, USA
Carl Kesselman, USC/ISI, Marina del Ray, CA, USA
Veronika Nefedova, Argonne National Laboratory, Argonne, IL, USA
Line Pouchard, Oak Ridge National Laboratory, Oak Ridge, TN, USA
Arie Shoshani, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
Alex Sim, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
Gary Strand, NCAR, Boulder, CO, USA
Dean Williams, Lawrence Livermore National Laboratory, Livermore, CA, USA

Global coupled Earth System models are vital tools for understanding potential future changes in our climate. As we move towards mid-decade, we will see new model realizations with higher grid resolution and the integration of many additional complex processes. The U.S. Department of Energy (DOE) is supporting an advanced climate simulation program that is aimed at accelerating the execution of climate models one hundred-fold by 2005 relative to the execution rate of today. This program, and other similar modeling and observational programs, are producing terabytes of data today and will produce petabytes in the future. This tremendous volume of data has the potential to revolutionize our understanding of our global Earth System. In order for this potential to be realized, geographically distributed teams of researchers must be able to manage and effectively and rapidly develop new knowledge from these massive, distributed data holdings and share the results with a broad community of other researchers, assessment groups, policy makers, and educators.

The Earth System Grid II (ESG-II), sponsored by the U.S. Dept. of Energy's Scientific Discovery Through Advanced Computing (SciDAC) program, is aimed at addressing this important challenge. The broad goal is to develop next generation tools that harness the combined potential of massive distributed data resources, remote computation, and high-bandwidth wide-area networks as an integrated resource for the research scientist. This integrative project spans a variety of technologies including Grid and DataGrid technology, the Globus Toolkit™, security infrastructure, OPeNDAP, metadata services, and climate analysis environments. In this presentation we will discuss goals, technical challenges, and emerging relationships with other related projects worldwide.

4. Geoscientific Data Amalgamation: A Computer Science Approach
N. L. Mohan, Osmania University, India

Rapidly changing environment of data, information and knowledge and their communication and exchange has created multidimensional opportunity for researcher that it could (i) avoid duplication of work, increases - (ii) the competitive spirit, (iii) the high quality of research and (iv) the interdisciplinary areas of research. In this context amalgamation of geo-scientific data assumes greater importance. The amalgamation and availability of numerical form of geo-scientific data are possible based on fundamental premise that it is open to public domain and quality of data is assured.

Availability of data pertaining to earth related sciences in general and geophysical data in particular can be classified into three categories -(a) Raw data form, (b) Processed or filtered form of raw data and (c) Theoretically computed form. Further, data are of two types - Static mode and Dynamic mode as certain type of data pertaining to exploration geophysical methods like gravity, magnetic, electrical, seismic, well logging etc are Static type. That is once the data is acquired it may remain static. On the contrary, the data pertaining to earth tides, geomagnetic field records, earthquake seismograms, Satellite data etc are of Dynamic type. One more dimension of data availability is Model Construction, based on Expert System Shells- an Artificial Intelligence approach. The Final version of data availability is through two formats- Graphical and Image forms.

It is not the question to make available numerical data but how to organize, manage, and update are challenging aspects in Earth-related sciences. Unlike other fields of science and engineering, geo-scientific data management needs altogether special approach. The author believes that geo-scientific community may need to understand certain important areas of computer science so that they could guide computer specialists appropriately to manage data efficiently and communicate to his earth science community effectively, in 2-D, 3-D numerical form, including graphical and imagery forms.

Data Base System is a computerized record-keeping system. Several operations like Adding data to new and empty files, Inserting data into existing files, Retrieving data from existing files, Changing data in existing files, Deleting data from existing files and Removing existing files are involved in record keeping system. Further, Data Base Architecture comprises three levels- (i) Internal Level, a centralized storage system as it would help to store all types of geo-scientific data at one place; (ii) Conceptual level, a specific type of data storage system as it would help to track the data by specialist in the concerned specialized area like earthquake data and (iii) External level, a user node connectivities as simultaneously number of users can track different types of geo-scientific data according to ones own interest. Also, from another angle an abstract view of the data can be broadly segmented into three levels - (a) the physical level gives the idea how the data are actually stored; (b) logical level indicates that what data are stored in the data base, and what relationships exist among those data and (c) view level represents only a part of entire data base.

One of the most important data bases, particularly form the point of geo-scientific data organization, is object oriented data system that would play a predominant role to organize several different sets of data and could refer to each other for better understanding, modeling and refining the models and making meaningful and correct inferences etc. The object oriented data arrangement is based on concepts like (a) inheritance, (b) polymorphism, (c) multiple inheritance etc. These respective concepts would help to inherit certain properties from parent entity apart from its own; a function with same name can perform different tasks by taking different sets of data; and a class or set of different classes can inherit different properties from several classes.

Parallel and distributed data bases do play important but limited roles in certain contexts of geo-scientific data amalgamation systems. That is, these data base approaches may help within a geo-scientific organization where certain confidentiality is required and does not want to throw on public domain initially for some time.

Artificial Intelligence is the most promising area that geo-scientific community should look into for data organization, modeling, searching, multi-dimensional construction and view of graphical and image models and data management etc. In view of data communication through high speed, wide band and wireless inter-net domain

systems, Knowledge Base and expert System Shells dominating the scene. Artificial Intelligence envelopes several areas like, data bases, object oriented representation, search and control strategies, matching techniques, knowledge organization and management, pattern recognition, visual image understanding, expert system architecture, machine learning, several types of learning techniques that include neural networks, problem solving methods, robotics, semantic nets, frames, cognitive modeling, data compression techniques etc.

Another important area is algorithmic design and analysis which is vital for geo-scientific data organization and management. It is very much important that algorithms must be designed based on how the geo-scientific data need to be arranged, such that search, retrieval, modification, insertion may be made user friendly.

Finally, the geo-scientific data amalgamation would be successful only if proper indexing, security, integraty and standardization or bench markings are taken care of.

---

5. The US National Virtual Observatory: Developing Information Infrastructure in Astronomy

A. Szalay, Johns Hopkins Univ, USA
D. De Young, Nat'l Optical Astronomy Obs, USA
R. Hanisch, Space Telescope Science Inst, USA
G. Helou, Cal Tech/IPAC, USA
R. Moore, San Diego Supercomputer Center, USA
E. Schreier, Space Telescope Science Inst, USA
R. Williams, Cal Tech/CACR, USA

The Virtual Observatory (VO) concept is rapidly becoming mature as a result of intensive activity in several countries. The VO will provide interoperability among very large and widely dispersed datasets, using current developments in computational and grid technology. As a result, the VO will open new frontiers of scientific investigation and public education through world-wide access to astronomical data sets and analysis tools. This paper will provide and overview of present VO activities in the US, together with a brief description of the future implementations of USNVO capabilities.

Track III-C-6: Data Centers

Chair: David Clark, NOAA National Geophysical Data Center, USA

1. Database Management at Indian Oceanographic Data Centre
Pravin D. Kunte, National Institute of Oceanography, India

Ocean studies are inherently interdisciplinary and therefore call for a controlled and integrated approach for information generation, processing and decision-making. In this context, Indian Oceanographic Data Centre (IODC) was established at National Institute of Oceanography (NIO) as early as 1964 for Institutional data management. Later IODC has been recognized as National Marine Data Centre (NMDC) in 1990 by Department of Ocean Development (DOD) as a national facility and as National Oceanographic Data Centre (NODC) during 1994 and as Responsible National Oceanographic Data Centre (RNODC) during 1998 as an International facility by Intergovernmental Commission (IOC, Paris) to acquire, archive entire spectrum of marine and oceanographic data and disseminate it to researchers, scientists and decision makers and other end users in form of value added product.

IODC acquires, processes, stores and disseminates the oceanographic data pertaining to Arabian Sea, Bay of Bengal, laccadive sea, Andaman & Nicobar sea and the Indian Ocean and share data with Indian Ocean rim countries. Voluminous data is managed using INGRES database management system on UNIX workstation and is preferably disseminated on CDs along with user-friendly searching and retrieving software. Efficient bibliographic information system, which forms an integral part of oceanographic Information system, is also maintained at NIO. Strong team of approximately 200 scientists working at NIO in different oceanographic disciplines support IODC activities. This manuscript concentrates on describing the organization and structure status, role and functions of IODC, management and quality control procedures, IODC's impact and future outlook. Finally based on 35 years of data management experience, important suggestions are incorporated.

2. CODATA in Africa "The Nigeran Data Program"
Kingsley Oise Momodu, Chairman CODATA Nigeria, Faculty of Dentistry, University of Benin, Nigeria

Approval was granted for Nigeria's membership into CODATA International in May 1998, in response to an application by the Federal Ministry of Science and Technology. The Nigerian CODATA committee has as its mandate, the task of providing liason between the scientifc community in Nigeria and the International Scientific Community. The Nigerian CODATA committee participated in the fourth International Ministerial meeting of the United Nationas Economic Commission for Africa (ECA) on development information on Thursday 23rd November 200 at the National Planning Commission conference room in Abuja, Nigeria. At the meeting, CODATA made the following observations:

1. That new research projects tend to get much more attention than already completed ones.
2. The continued processing of data from old projects through secondary analysis is often neglected.
3. A lack of directories that describes what data sets exists, where they are located and how users can access them, which leads to unnecessary dublication of efforts.
4. Lack of a viable network among scientists.
5. That the existence of data is unknown outside the original scientific group or agencies that generated them and even if known, information is not provided for a potential user to access their relevance.
6. That scientists in Africa are fundamentally poorly paid.

The Nigerian CODATA is making spirited efforts to respond adequately to the necessities for preserving data by establishing a data program which represents a strategy for the compilation and dissemination of scientific data. This program will help the local Scientific community in Nigeria take adventage of the opportunities and expertise offered by CODATA International which has made a commitment to assist Scientific research institutes and the local Scientific community in the area of database development.

This initaitive is also designed to complement the activities of the task group on reliable scientific data sources in Africa. The scientific activity for year 2002 is a survey and cataloguing of potential data sources which will be web-based.

---

3. The 'Centre de Données de la Physique des Plasmas' (CDPP, Plasma Physics Data Centre), a new generation of Data Centre
M. Nonon-Latapie, C.C. Harvey, Centre National d'Etudes Spatiales, France

The CDPP results from a joint initiative of the CNRS (Centre National de la Recherche Scientifique) and the CNES (Centre National d'Etudes Spatiales). Its principal objectives are to ensure the long term preservation of data relevant to the physics of naturally occurring plasmas, to render this data easily accessible, and to encourage its analysis. The data is produced by instruments, in space or on the ground, which study the ionised regions of space near the Earth and elsewhere in the solar system.

The principal users of this data centre are space scientists, wherever they are located. This data centre is located in Toulouse (France), and it uses a computer system which is accessible via the Internet (http://cdpp.cesr.fr/english/index.html). This system offers several services : firstly the possibility to search for and retrieve scientific data, but also access to the "metadata" archived in association with this data, as the relevant documentation and quicklook data (graphical representations). Several tools are available to help the user to search for data. The CDPP has been accessible since October 1999. Since then its data holding and the services offered have been steadily augmented and developed.

After a brief presentation of the objectives, the organisation, and the services currently offered by the CDPP, this paper will concentrate on :

- the system architecture (based on the ISO "Reference Model for an Open Archival Information System")
- the data model
- the standards used to format and describe the archived data
- the crucial operation of ingesting new data; this function is based on the description of all delivered data entities via a dictionary.

Operational experience and new technical developments now being studied will also be presented.

4. A Space Physics Archive Search Engine (SPASE) for Data Finding, Comparison, and Retrieval
James R. Thieman, National Space Science Data Center, NASA/GSFC, USA
Stephen Hughes and Daniel Crichton, NASA Jet Propulsion Laboratory, USA

The diversity and volume of space physics data available electronicallyhas become so great that it is presently impossible to keep track of what information exists from a particular time or region of space. With current technology (especially the World Wide Web - WWW) it is possible to provide an easy way to determine the existence and location of data of interest via queries to network services with a relatively simple user interface. An international group of space physics data centers is developing such an interface system, called the Space Physics Archive Search Engine (SPASE). Space physicists have a wealth of network-based research tools available to them, including mission- and facility-based data archive and catalogue services (with great depth of information for some projects). Many comprehensive lists of URL's have been put together to provide a minimal search capability for data. One recent effort to gather a list of data sources resulted in an assembly of nearly 100 URL's and many important archives had still been missed. These lists are difficult to maintain and change constantly. However, even with these lists it is not possible to ask a simple question such as "where can I find observations in the polar cusp in 1993?" without doing extensive, manual searches on separate data services.

The only hope for a comprehensive, automated search service is to have data centers/archives make their own information available to other data centers and to users in a manner that will facilitate multiarchive searching. Nearly all space physics data providers have WWW services that allow at least a basic search capability, and many also provide more specialized interfaces that support complex queries and/or complex data structures, but each of these services is different. The SPASE effort is creating a simple, XML-based common search capability and a common data dictionary that would allow users to search all participating archives with topics and time frames such as "polar cusp" and "the year 1993". The result would be a list of archives with relevant data. More advanced services at later stages of the project would allow intercomparison of search results to find, for example, overlapping data intervals. Retrieval of the relevant data sets or parts of the data sets would also be supported. The first stages of the project are based on the application of Object Oriented Data Technology (OODT - see http://oodt.jpl.nasa.gov/about.html) to the cross archive search capability. The initial effort also includes the derivation of a common data dictionary for facilitating the searches. The current state of these efforts and plans for the future will be reviewed.

Track III-D-5: Information Management Systems

Chair: Glen Newton, CISTI, National Research Council of Canada, Canada

The efficient and effective collection and management of data, information and knowledge is becoming more difficult, due to the volume and complexity of this information. Greater demands on system architectures, system design, networks and protocols are the catalyst for innovative solutions in management of information.

Applications and systems being researched and developed capture dimensions of the various issues presented to the community, and represent aspects of new paradigms for future solutions.

Some of the areas to be examined include:

- Systems for Coupling and Integrating Heterogeneous Data Sources
- Web services for data
- Decision Support Systems
- Intelligent Agents, Multi-Agent Systems, Agent-Oriented Programming
- Interactive and Multimedia Web Applications
- Internet and Collaborative Computing
- Multimedia Database Applications

1. Informatics Based Design Of Materials
Krishna Rajan, Rensselaer Polytechnic Institute, USA

In this presentation we demonstrate the use of a variety of data mining tools for both classification and prediction of materials properties. Specific applications of a variety of multivariate analysis techniques are discussed. The use of such tools has to be coupled to a fundamental understanding of the physics and chemistry of the materials science issues. In this talk we demonstrate the use of informatics strategies with examples including the design of new semiconductor alloys and how we can extend the concept of bandgap engineering to the development of "virtual" materials. The use of the combination of these approaches when integrated with the correct types of descriptors, allows informatics methodologies to be a powerful computational methodology for materials design.

2. XML-Based Factual Databases: A Case Study of Insect and Terrestrial Arthropod Animals
Taehee Kim Ph.D., School of Multimedia Engineering, Youngsan University, South Korea
Kang-Hyuk Lee, Ph.D., Department of Multimedia Engineering, Tongmyung University of Information Technology, South Korea

XML (eXtensible Markup Language) serves as the de facto standard for document exchange in many data applications and information technologies. Its application areas span from ecommerce to mobile communication. An XML document describes not only the data structure, but also the document semantics. Thus, a domain specific, and yet self-contained document could successfully be built by using XML. Building and servicing factual databases in the XML format could provide such benefits as easy data exchange, economic data abstraction, and thin interface to other XML applications like e-commerce.

This paper reports an implementation of factual databases and thier service in terms of XML technologies. The database of insect and terrestrial arthropod animals has been constructed as an example. The insect database contains

the intrinsic information on characteristics of Korean insects while the terrestrial arthropod animal database contains the related bibliographical information. Data Types and document structures were implemented. Data type definitions (DTDs) were then implemented for data validation. Microsoft SQL Server incorporated with Active Service Pages was used as our implementation framework. A web database service had also been built.

Based on the implementation, this paper then discusses issues related to XML-based factual databases. We emphasize that document design ought to be carried out in order to achieve maximal compatibility and scalability. We then point out that an information service system could better be built by exploiting the self-describing characteristics of XML documents.

---

3. A comprehensive and efficient "OAIS compliant" data center based on standardized XML technologies
Thierry Levoir and Marco Freschi, Centre National d'Etudes Spatiales, France

The OAIS (Open Archival Information System) Reference Model provides a framework to create an archive (consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community). It offers also a real help to design: ingest, data management, administration and data access systems. XML stands for eXtensible Markup Language. XML starts as a way to mark up content, but it soon became clear that XML also provided a way to describe structured and semi-structured data thus making the usage as a data storage and interchange format. Many related languages, formats, technologies like SOAP, XML Query, XML-RPC, WSDL, Schema, ... are still coming to provide solutions to almost all problems!

With such technologies, we can define many different architectures. Due to the vastness of the problem, it is quite difficult to describe all the possible solutions, so the article is intended to describe a possible architecture of a system where, the organization of data and their usage, is defined in accordance with the OAIS reference model. The article takes cue on the needed to update an existing data center, providing some features like platform independence, human readable format of data and easy extensibility for new type of data. All these advantages seem to be supplied by XML and Java. XML and Java together can certainly be used to create some very interesting applications from application servers to better searchable web sites. It also offers an easy and efficient way to be interoperable.

However, it is sometimes very difficult to understand where everything really fits. The article attempts to clarify the role of each single object inside of a data center, providing as result, the complete description of a system including its architecture. A section of the article is also dedicated to the problem to make data persistent on the database, the choice of this support often involves an automatic choice of a query language to retrieve data from the database and a strategy to store them.

---

4. XML-based Metadata Management for INPA's Biological Data
J. L. Campos dos Santos, International Institute for Geo-Information Science and Earth Observation - ITC, The Netherlands and The National Institute for Amazon Research - INPA, Brazil
R. A. de By, International Institute for Geo-Information Science and Earth Observation - ITC, The Netherlands

For more than a century, Amazonian biological data have been collected, primarily by single or small group of researchers in small areas over relatively short periods of time. Questions on "how ecological patterns and processes vary in time and space, and what are the causes and consequences of this variability" are still in questioning. For such questions to be properly answered, far more documented data are required than could feasibly be collected, managed, and analysed in a single organisation. Since biological data sets are neither perfect nor intuitive, they are shared in a close range to the data producers, who know the subject. Few additional information are needed for data sets to be used and interpreted. Research teams outside of the specific subject area need highly detailed documentation to accurately interpret and analyse historic or long-term data sets, as well as, data from complex experiments.

Usually, researchers refer to their data as raw data, which are structured in rows and columns of numeric or encoded sampling observations. The usefulness of such data can only be assessed when they are associated to either a theoretical or conceptual model. This requires understanding of the type of variable, the units adopted, potential biases in the measurement, sampling methodology and a series of facts that are not represented in the raw data, but rather in the metadata. Data and metadata combined within a conceptual framework produces the so needed information. Additionally, information can be lost through degradation of the raw data or lack of metadata. The loss of metadata can occur throughout the period of data collection and the rate of loss can increase after the results of the research have been published or the experiment ends. Specific details are most likely to be lost due to the abandonment of data forms and field notes. Metadata will ensure to data users the ability to locate and understand data through time.

This paper presents an XML-based solution for the management of metadata biological profiles via the Web. We have adopted the FGDC Metadata Standard, which incorporates the Biological Data Profile, and is represented as an XML schema. The schema is mapped to a well-formed biological metadata template. The XML metadata template can be deployed to users together with an XML Editor. The editor uploads the XML file and allows them to insert all the metadata information. After this process, the biological metadata can be submitted for certification and stored in an XML repository. The repository accepts a large number of well-formed XML metadata and maintains a single data representation of all the files it receives. The metadata can be retrieved, updated, or removed from the repository that once it is indexed, search and query are available. This solution is in test at the National Institute for Amazon Research (INPA) within the Biological Collection Program.

---

5. Incorporation of Meta-Data in Content Management & e-Learning

Horst Bögel, Robert Spiske and Thurid Moenke, Department of Chemistry of the Martin-Luther-University Halle-Wittenberg, Germany

In nearly all scientific disciplines experiments, observations or computer simulations produce more and more data. Those data have to be stored, archived and retrieved for inspection or later re-use. In this context the inclusion of 'Meta-Data' becomes important to have access to certain data and pieces of information.

There are three developments recently to be taken into account:

- Content Management Systems (CMS) are used for teamwork and 'timeline' co-operation
- use of eXtended Hypertext Markup Language (XML) to separate the content from the layout for presentation (DTD, XSL, XPATH, XSLT)
- online Learning using WEB-technology develops for a powerful multimedia-based education system

All those key-processes are based on huge amount of date and the relations between them (this can be called information).

If we want to keep pace with necessities, we have to develop and use these new techniques for making progress.

We report about some tools (written in Java) for handling of data and the development of a unique WWW-Based Learning five-years project (BMBF - Geman Federal Ministry for Education and Research) in chemistry, to incorporate data (3D-structures, spectra, properties) and their visualizations in order to favour a research-oriented way of learning. The students have access to computational methods in the network, to carry out open-end calculations using different methods (e.g. semi-empirical and ab initio MO calculations to generate the electronic structure and the orbitals of molecules).

Track IV-A-3: Spatial Data Issues

Chair: Harlan Onsrud, University of Maine, USA

1. Spatio-Temporal Data Visualization for EOS Data
Udeepta Bordoloi, Ohio State University, USA
Han-Wei Shen, Ohio State University, USA
David Kao, NASA, USA

Visualizing high dimensional data has been traditionally a difficult problem. The data domain can be 2D or 3D, and the data itself can be scalar, vector, or higher dimensional. For example, samples of probability density function defined in a volume with time-varying characteristics would give us a five-dimensional dataset. Coupled with the enormously large amounts of data samples many applications now produce, effective analysis of these high dimensional data becomes very challenging.

In many applications (e.g. meteorological studies), the behavior of a variable over time is of prime importance. Scientists studying spatial maps of satellite data often need to compare and understand how the maps change with time. In case of probability densities, the extra dimension makes the job even more difficult. One method to display the time-varying properties of the data is to play an animation. For density data, we can animate an appropriate statistical summary (e.g., the mean) of the probability densities. However, this is not the best option in many cases. Consider the following question: "How does the map change from one time-step to the next near the start of the animation, and how does it change near the end of the animation?" A comparison of the temporal change near the beginning of the animation to the temporal change near the end of the animation is better performed if the user can see some representation of the change in one snap-shot, rather than having to deduce the change indirectly by following a movie. Moreover, animation is effective only if the user can visually track the changes/movements in the animating visuals. For example, sudden changes in values cannot be quantitatively appreciated by the user. They can also result in popping artifacts. Another situation in which movies are not effective is when a large number of regions in the maps move simultaneously. The user is easily overwhelmed trying to follow all the movements.

In this paper, we present a method to visualize the behavior of 2D scalar and probability density function data in time. Our goal is to present the user with information which helps him/her in appreciating the temporal behavior of a given region in the 2D domain. The multi-resolution nature of the algorithm allows the user to dynamically choose the visualization at various levels of detail in both spatial and temporal domains. The visualization technique consists of two main components. First, a hierarchical spatio-temporal clustering algorithm is used to divide the map into disjoint spatio-temporal regions. The clustering results for a single time step give a snapshot of spatial information, while animation allows us to track how the clusters move and change their shapes with time. Second, a modified theme-river technique is used in our method to display the temporal variation of any property associated with the data at different levels of detail, either for static or changing regions. We present examples of NDVI data representing monthly global vegetation cover, and of probability density data generated from ocean modeling.

2. Spatio-Temporal Database Support for Long-Range Scientific Data
Martin Breunig, Institute of Environmental Sciences, University of Vechta, Germany
Serge Shumilov, Institute of Computer Science III, University of Bonn, Germany

Hitherto, database support for spatio-temporal applications is not yet part of standard DBMS. However, applications like telematics, navigation systems, medicine, geology, and others require database queries referring to the location of moving objects. In real time navigation systems, the position of large sets of moving cars has to be determined

within seconds. In patient-based medical computer systems, to take another example, the relevant progress of diseases has to be examined during days, weeks or even years.

Finally, geological processes like the backward restoration of basins include time intervals of several 1000 years to be considered between every documented snapshot of the database. We restrict ourselves to the requirements of long-range applications like the simulation of geological processes. An example for a spatio-temporal database service in geology is given. The two components of this service provide version management and the temporal integrity checking of geo-objects, respectively. In the given example, the location change of a 3D moving object O(t) between two time steps ti and ti+1 may have three reasons: the location change of the partial or complete geometry of O(t) at time ti caused by geometric mappings like translation, rotation etc., the change of the scale of O (t) at time ti caused by zooming (change of the size of the object) or the change of the shape of O(t) at time ti caused by mappings of one or more single points of its geometry.

These three reasons can also occur in combination with each other. Furthermore, 3D moving objects may be decomposed into several components and later sequently merge to a single object again. The relevant database operations needed to map the objects into the database (compose and merge) give a mapping between the IDs of all objects between two directly following time steps. We show that the simulation of long-range processes can be effectively supported by set-oriented spatio-temporal database operations. Among other aspects, this leads to a better understanding of the history of geological rock formations. The specifications of the operations are given in C++ program code. We describe how the proposed spatio-temporal operations are integrated into GeoToolKit, an object-oriented database kernel system developed for the support of 3D/4D applications. In our future work we intend to evaluate the presented spatio-temporal database operations in benchmarks with large data sets within the open GeoToolKit system architecture as part of a public database service for spatio-temporal applications.

---

3. Web Visualization for Spatio-Temporal Referenced Multimedia Data

Paule-Annick Davoine and Hervé Martin, Laboratoire LSR-IMAG, Equipe SIGMA, France

Web and multimedia technologies enhance possibilities to develop software for managing, displaying and distributing spatially-referenced information. A lot of works deal with Internet based cartographic visualization. More and more geographic applications have to integrate both a temporal and a multimedia dimension. This kind of information appears more complex than usual spatial information linked with statistical data such as economical, demographic or ecological data. The main problem for Geographical Information Systems (GIS) is to merge qualitative and multimedia information with information related to time and space. Moreover, spatial and temporal references may be heterogeneous and discontinuous. Actually, current GIS are not suited to the use and to the visualization of this kind of information. In this paper, we show how multimedia web information systems may be used to model and to navigate across spatio-temporal referenced multimedia data. This work has been realized in the framework of an European project, named SPHERE, on historical natural hazards.

In a first time we explain how Unified Modelling Language (UML) language allows to take into account various user requirements. We focus on two important features of GIS: how to specify system functionalities and how to capture spatio-temporal referenced multimedia data. To illustrate the former point, we present the main functionalities of the web visualisation interface that we have implemented during the SPHERE project. We also explain the technological choices used to develop this tool. We have developed Java tool based on a client-server architecture including an Apache Web server and a client browser for running Java applets. This software allows to navigate across information space according to spatial and temporal features and to visualize simultaneously and interactively cartographic, temporal and documentary aspects of information.

Keywords: Web and Database, Geographical Information System (GIS), information system, multimedia, spatio-temporal referenced information, UML modelling, visualisation interface.

Track IV-A-5: Information Infrastructure for Science and Technology

Horst Kremers, Eng., Comp. Sci., Berlin, Germany

The manageability of complex information systems for multidisciplinary cooperation and its use in decision support depends on basic methods and techniques that cover application layers, such as:

- the role of basic sets of information in global and national information infrastructures
- access, compatibility, and interoperability
- documentation of information models
- validation procedures and quality control
- financial and legal aspects (including copyright)
- enabling cooperation on information
- archiving
- education

This session offers opportunities to present best practices in information infrastructure, as well as discussing the methodological backgrounds and potential ways to support the creation of national and global interdisciplinary information infrastructures. The session, of course, has cross-links to other sessions in the CODATA conference. Topics here are to be discussed in their strategic importance with respect to enabling freedom of information, as well as enabling reliable communication and cooperation in the information society.

1. Exchange Of Heterogeneous Information Concepts And Systems

Hélène Bestougeff, CODATA - France
Jacques-Emile Dubois, ITODYS, Université de Paris VII - France and Past-President, CODATA

Today, especially with the development of networking and the internet, the exchange of heterogeneous information leading towards better interdisciplinary co-operation is a vital issue. In this framework, several technical and organizational problems must be solved. Integration deals with developing architectures and frameworks as well as techniques for integrating schemas and data. Different approaches to interrelate the source systems and user's queries are possible depending on the degree of coupling between the original data sources and the resulting system.

However, integration is just a first essential step towards more sophisticated architectures which are developed towards management decisions support. These architectures, grouped under the term of Data Warehouses are subject oriented and involve the integration of current and historical data.

The third aspect of heterogeneous information exchange is knowledge management, information mining systems, and web information management. The web is now part of almost all organizations. Therefore, the databases and the warehouses of specific networks endowed with adequate metadata have to operate on the web. Moreover, the management of unstructured and multimedia data such as text, images, audio and video presents new original challenges.

This paper will present, in a systematic way, concepts and systems dealing with these problems and drawing on particular results and examples from the just published book by Kluwer " Heterogeneous Information Exchange and Organizational Hubs".(H.Bestougeff, J.E. Dubois, B.Thuraisingham, Editors) containing 15 original chapters covering:

I Heterogeneous Database Integration: Concepts and Strategies
II Data warehousing: Models and Architectures
III Sharing Information and Knowledge.

2. Information Infrastructure: The Dynamic Mosaic
Horst Kremers, Eng., Comp. Sci., Berlin, Germany

The various aspects of Information Infrastructure presented and dicussed in this session and in contributions troughout this conference give an overview of the specific role of CODATA shaping and developing this field in its specific competence and interest. The mosaic that this conference shows can be completed as well as it can be clearly distinguished from from other activities in Information Infrastructure at national and at international level. In addition, the various contributions have shown that this field is under dynamic development. This allows the discussion of a potential CODATA strategic position and of actions that would promote this development because of its growing relevance in an appropriate way.

3. Cooperative Canadian/US Project: An Experiment in Sharing Geospatial Data Cross-Border
Milo Robinson, US Federal Geographic Data Committee, USA
Marc LeMaire, Mapping Services Branch, USA

The US Federal Geographic Data Committee (FGDC) and its Canadian counterpart, GeoConnections, have developed several cooperative projects to develop a common spatial data infrastructure. To better understand the challenges and complexities of transboundary spatial data issues, GeoConnections and the Federal Geographic Data Committee jointly funded two collaborative demonstration projects covering a common geographic project that crosses the border and addressing a common issue, that of sharing data with our neighbors. These collaborative projects cover the Red River Basin (Roseau River and Pembina River Basin) and the Yukon to Yellowstone (Crown of the Continent Study Area). The results of these international spatial data demonstration projects, as well as new joint activities, will be discussed.

4. Current Trends in the Global Spatial Data Infrastructure: Evolution from National to Global Focus
Alan R. Stevens, Global Spatial Data Infrastructure (GSDI) Secretariat, USA

In the late 1980's many organizations from state, local and tribal governments, the academic community, and the private sector within the United States came together to encourage common practices and uniform standards in digital geographic (map) data collection, processing, archiving, and sharing. The National Spatial Data Infrastructure (NSDI) encompasses policies, standards, and procedures for organizations to cooperatively produce and share georeferenced information and data. The major emphasis now has turned toward Geospatial One-stop initiatives, Implementation Teams (I-Teams), and Homeland Security for better governance, but all still aimed at facilitating the building of the NSDI.

In the mid '90s other nations began to recognize that tremendous efficiencies and cost savings could be realized by reducing duplicative data collection, procession, archive and distribution not only within their own borders but across international boundaries as well. A small group, at first, spawned what is now known as the Global Spatial Data Infrastructure (GSDI). This group now has grown to over 40 nations and consists of government agencies, NGO's, academic institutions, other global initiatives, and a significant contingent of the private sector in the geosopatial industry. Industry is excited to be involved because they realize that common standards will increase demand for data from domestic customers and will expand the awareness within emerging nations further increasing the client base. The GSDI has incorporated as a non-profit organization so it can partner with others in securing funds to encourage and accelerate the development of National and Regional Spatial Data Infrastructures in fledgling organizations and countries.

Track IV-B-3: Data Portals

Chair: Juncai Ma, Institute of Microbiology, Chinese Academy of Sciences, China

1. Information Society Technologies Promotion for New Independent States
A.D. Gvishiani, Director of the Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia
J. Babot, Head of the E Work Sector in European Commission, Belguim
J.Bonnin, Institut de Physique du Globe de Strasbourg, France

Recently proposed cluster project CLUSTER-PRO will unite and coordinate the activities of the five Information Society Technology (IST) program projects now running by the European Commission in Baltic (TELEBALT), Eastern European (E3WORK and TEAMwork) and CIS (WISTCIS and TELESOL) countries in order to promote new information technologies for scientific and technological data handling in these countries. French Committee on Data for Science and Technology (CODATA FRANCE) serves as the coordinator of the proposed CLUSTER-PRO project.

In the presentation, all the five projects under clustering will be described. The goals of these projects are to promote modern teleworking tools for scientific, technological and business data handling and exchange between EU member states and EU pre-accession and third countries. Main goal of the cluster is to create a common structure of concentration between the existing projects, with a common portal, cross exchange and adaptation of results, common action plan for dissemination. One of the objectives is the elaboration of the cross-project Web sites that will focus on new opportunities of teleworking in scientific and technological data acquisition and exchange. Education, research, business, tele-medicine, new employment opportunities promotion and environmental protection in all participating countries are among the cluster project objectives. Another objective is the cross-project training actions using the courses and e-learning systems developed by the five clustered projects, which will be adopted for the whole range of countries. Cross-project training actions will be implemented in face-to-face mode at CLUSTERPRO and the clustered projects gatherings, and virtually through the Web sites. An information portal "New opportunities for EU-CLUSTER-PRO countries teleworking" will be developed.

2. Data, Information, and Knowledge Management of a Solar UV Data Network for Coating Materials
Lawrence J. Kaetzel, K-Systems, Prescott, Arizona, USA and
Jonathan W. Martin, National Institute of Standards and Technology, Gaithersburg, MD, USA

A major factor in understanding the performance of coating materials and products requires the use of systematic methods for acquiring, recording, and interpreting laboratory and field test results and environmental conditions. The verified data from these sources can then be used with computer-based models for more accurately predicting materials performance. A solar ultra-violet data network has been created by the National Institute of Standards and Technology, Gaithersburg, Maryland. The network measures, collects and archives weather measurements for use in predicting the performance of automotive and architectural coatings. Operation of the network is performed as a collaborative effort among several U.S. Government agencies and private industry organizations. The network currently consists of 8 field locations operating in the United States that are equipped with solar spectroradiometers and weather stations. Data from the network is evaluated and stored electronically, then used in scientific analysis as applied to materials performance and biological studies.

This paper presents the efforts to ensure data integrity; the methodologies used to represent the data, information, and knowledge in a consistent manner; and the computer-based methods (e.g., computer-based models, decision-support or smart modules) developed to assist the knowledge consumer in determining relevance and to assist in the interpretation of the measurements. The paper will first discuss the application of the data as applied to coating performance and its use with computer-based modules, followed by a discussion of knowledge management methods.

---

3. Development of Information Infrastructure for Microbiological Study in China

Juncai MA, Pengtao Liu, Yoshihiro Ichiyanagi and Yimei Lao
Institute of Microbiology, Chinese Academy of Sciences, China

China is the country with rich microbial diversity. Chinese have long history to use microorganism in industry, agriculture and medicine. Microbial Information Network of China(MICRO-NET) was founded in 1994. Its purposes are to collect and share microbial information resources in China, and to introduce the popular foreign information resources to Chinese scientists. MICRO-NET now has become a important basic information infrastructure for the daily microbiological study in China.

1. Information System for National Culture Collections
In China, there are two main type culture collections, CCCCM and CTCCCAS. In CCCCM (China Committee of Culture Collections of Microorganisms), there are 12 national culture collections divided into general microbiological, agricultural, industrial, medical, antibiotic, veterinary and forestry center. China Catalogue of Cultures (English version) includes 10,716 strains of bacteria, actinomyces, yeast, fungi and virus. The CTCCCAS (Committee on Type Culture Collection of CAS) was organized in April, 1996. There are seven Culture banks under CTCCCAS, namely, Microbial Culture Bank, Cell Bank, Gene Bank, Virus Bank, Kunming Cell Bank and Freshwater Algae Bank, the Rare, Endangered and Endemic Plant Germplasm Bank, Marine Biological Germplasm Storehouse and In Vitro Plant Germplasm Collection. Up to the present, 15,418 cultures have been collected. We have developed the online information systems for CCCCM and CTCCCAS, now all the culture holds in the two systems can be searched through Internet.

2. Electronic version of Flora Fungorum Sinicorum (full text)
In the coming five years, about one hundred books of FFS will be published. The Electronic Version for full text of Flora Fungorum Sinicorum is supported by NSFC(National Natural Science Foundation of China). Up to now, full text of 10 books has been inputted into database.

3. China Node of International Bio-mirror Network
Bio-Mirror is a project under APAN (Asia Pacific Advanced Network, http://www.apan.net). The China Node of International Bio-Mirror Network is set up in our center. We started to mirror some important international databases in 1997. Now more than 30 databases including International Nucleotide Sequence Database (DDBJ/EMBL/GENBANK) are mirrored. By SRS, a online full text search function is available for all the databases.

4. Integration Information Search Engine
Network has become an indispensable tool for daily research study. There are huge number of biological species information in the Internet, such as catalogue information, culture collection information, literature information, genetic information and so on. We are developing a meta search engine named SPECIES-Info to help biologists find species information in Internet more easily and more efficiently. In SPECIES-Info, we focus the information world wide on culture collection, online journals, genetic data, patents and so on. Now XML has been used for SPECIES-Info.

4. Data Grid Structure of Scientific Databases of CAS

Kai Nan and Baoping Yan, Computer Network Information Center, Chinese Academy of Sciences, China

Data grid technology is an emerging technology to solve the challenging problems in data resource sharing, especially distributed and heterogeneous data. Although it is not well done, data grid has become a prominent technical trend and attracted more and more interests. Currently quite a few data grids are coming into being across the world. Scientific Databases (SDB) of CAS is a long-term project from 1983. There are a mass of scientific data in SDB, which are typically distributed and heterogeneous. So SDB can take greatadvantage of data grid technology. This paper gives a structural design of SDG (Scientific Data Grid), the application of data grid technology in SDB.

Data Science

Track I-D-5: Data Science

Chair: Jacques-Emile Dubois, ITODYS, Université de Paris VII - France and Past-President, CODATA

---

1. Quality Control of Data in Data-Sharing Practices and Regulations
Paul Wouters and Anne Beaulieu, Networked Research and Digital Information (Nerdi), NIWI-KNAW, The Royal Netherlands Academy of Arts and Sciences, The Netherlands

Scientific research is generating increasing amounts of data. Overall, in each year more data has been generated than in all years before combined. At the same time, knowledge production is becoming more dependent on data sets. This puts the question of quality control of the data center stage. How is the scientific system coping with the formidable task of controlling for the quality of this flood of data? One area in which this question has not yet been fully explored is the domain of data-sharing practices and regulations. The need to share data among researchers and between researchers and the public has been put on the agenda at the level of science policy (Franken 2000), partly out of fear that the system might not be able to cope with the abundance of data. Data sharing is not only a technical issue, but a complex social process in which researchers have to balance different pressures and tensions.

Basically, two different modes of data sharing can be distinguished: peer-to-peer forms of data sharing and repository-based data sharing. In the first mode, researchers communicate directly with each other. In the second mode, there is a distance between the supplier of data and the user in which the rules of the specific data repository determine the conditions of data sharing. In both modes, the existence or lack of trust between the data supplier and the data user is crucial, though in different configurations. If data sharing becomes increasingly mediated by information and communication technologies, and hence less dependent on face to face communication, the generation of trust will have to be organised differently (Wouters and Beaulieu 2001). The same holds for forms of quality control of the data. How do researchers check for the quality in peer to peer data sharing? And how have data repositories and archives taken care of the need for quality control of the data supplied? Which dimensions of social relationships seem to be crucial in data quality control? Which technical solutions have been embedded in this social process and what role has been played by information and communication technologies?

This paper addresses these questions in a number of different scientific fields (among others functional brain imaging, high energy physics, astronomy, and molecular biology) because different scientific fields tend to display different configurations of these social processes.

References:

H. Franken (2000), “Conference Conclusions” in: Access to Publicly Financed Research, The Global Research Village III Conference, Conference Report (P. Schröder, ed.), NIWI-KNAW, Amsterdam.

Paul Wouters and Anne Beaulieu (2001), Trust Building and Data Sharing - an exploration of research practices, technologies and policies. Research Project Proposal, OECD/CSTP Working Group on Datasharing.

2. Distributed Oriented Massive Data Management: Progressive Algorithms and Data Structures
Rita Borgo, Visual Computing Group, Consiglio Nazionale delle Ricerche (C.N.R.), Italy
Valerio Pascucci, Lawrence Livermore National Laboratory (LLNL), USA

Projects dealing with massive amounts of data need to carefully consider all aspects of data acquisition, storage, retrieval and navigation. The recent growth in size of large simulation datasets still surpasses the combined advances in hardware infrastructure and processing algorithms for scientific visualization. The cost of storing and visualizing such datasets is prohibitive, so that only one out of every hundred time-steps can be really stored and visualized.

As a consequence interactive visualization of results is going to become increasingly difficult, especially as a daily routine from a desktop. High frequency of I/O operations starts dominating the overall running time. The visualization stage of the modeling-simulation-analysis activity, still the ideal effective way for scientists to gain qualitative understanding simulations results, becomes then the bottleneck of the entire process. In this panorama the efficiency of a visualization algorithm must be evaluated in the context of end-to-end systems instead of being optimized individually. There is a need at system level to design the visualization process as a pipeline of modules able to process data in stages creating a flow of data that need themselves to be optimized globally with respect to magnitude and location of available resources. To address these issues we propose an elegant and simple to implement framework for performing out-ofcore visualization and view dependent refinement of large volume datasets. We adopt a method for view dependent refinement that relies on longest edge-bisection strategies yet introducing a new method for extending the technique to the field of Volume Visualization while keeping untouched the simplicity of the technique itself. Results in this field are applicable in parallel and distributed computing ranging from cluster of PC's to more complex and expensive architectures. In our work we present a new progressive visualization algorithm where the input grid is traversed and organized in a hierarchical structure (from coarse level to fine level) and subsequent levels of detail are constructed and displayed to improve the output image. We uncouple the data extraction from its display: the hierarchy is built by one process that traverses the input 3D mesh while a second process performs the traversal and display. The scheme allows us to render at any given time partial results while the computation of the complete hierarchy makes progress. The regularity of the hierarchy allows the creation of a good data-partitioning scheme that allows us to balance processing time and data migration time still maintaining simplicity and memory/computing efficiency.

3. Knowledge Management in Physicochemical Property Databases - Knowledge Recovery and Retrieval of NIST/TRC Source Data System
Qian Dong, Thermodynamics Research Center (TRC), National Institute of Standards and Technology (NIST), USA
Xinjian Yan, Robert D. Chirico, Randolph C. Wilhoit, Michael Frenkel

Knowledge management has become more and more important to physicochemical databases that are generally characterized by their complexity in terms of chemical system identifiers, sets of property values, the relevant state variables, estimates of uncertainty, and a variety of other metadata. The need for automation of database operation, for assurance of high data quality, and for the availability and accessibility of data sources and knowledge is a driving force toward knowledge management in the scientific database field. Nevertheless, current relational database technology makes the construction and maintenance of database systems of such kind tedious and error prone, and it provides less support than the development of physicochemical databases requires.

The NIST/TRC SOURCE data system is an extensive repository system of experimental thermophysical and thermochemical properties and relevant measurement information that have been reported in the world's scientific literature. It currently consists of nearly 2 million records for 30,000 chemicals including pure compounds, mixtures, and reaction systems, which have already created both a need and an opportunity for establishing a knowledge infrastructure and intelligent supporting systems for the core database. Every major stage of database operations and management, such as data structure design, data entry preparation, effective data quality assurance, as well as intelligent retrieval systems, depends to a degree on substantial domain knowledge. Domain knowledge regarding characteristics of compounds and properties, measurement methods, sample purity, estimation of uncertainties, data

range and condition, as well as property data consistency are automatically captured and then represented within the database. Based upon this solid knowledge infrastructure, intelligent supporting systems are being built to assist (1) complex data entry preparation, (2) effective data quality assurance, (3) best data and model recommendation, and (4) knowledge retrieval.

In brief, the NIST/TRC SOURCE data system is a three-tier architecture. The first tier is considered as a relational database management system, the second tier refers to knowledge infrastructure, and the last represents intelligent supporting systems consisting of computing algorithms, methods, and tools to carry out particular tasks of database development and maintenance. The goals of the latter two tiers are to realize the intelligent management of scientific databases based on relational model. The development of knowledge infrastructure and intelligent supporting systems is described in the presentation.

---

4. Multi-Aspect Evaluation of Data Quality in Scientific Databases
Juliusz L. Kulikowski, Institute of Biocybernetics and Biomedical Engineering c/o the Polish Academy of Sciences, Poland

The problem of data quality evaluation arises both, when a database is to be designed and when database customers are going to use data in investigations, learning and/or decision making. However, it is not quite clear what does it mean, exactly, that the quality of some given data is high or, even, it is higher than this of some other ones. Of course, it suggests that a data quality evaluation method is possible. If so, it should reflect the data utility value, but can it be based on a numerical quality scale? It was shown by the author (1982) that information utility value is a multi-component vector rather than a scalar.Its components should characterise such information features as its relevance, actuality, credibility, accuracy, completeness, acceptability, etc. Therefore, data quality evaluation should be based on vectors ordering concepts. For this purpose the Kantorovitsch's proposal of a semi-ordered linear space (K-space) can be used. In this case vector components should satisfy the general vector-algebra assumptions concerning additivity and multiplication by real numbers. This is possible if data quality features are defined in an adequate way. It is also desired that data quality evaluation is extended on data sets. In K-space this can be reached in several ways, by introduction of the notions of: 1/minimum guaranteed and maximum possible data quality, 2/ average data quality, 3/ median data quality. In general, the systems of single data quality and data quality sets evaluation are not identical.
For example, a notion of data set redundancy (being an important component of its quality evaluation) is not applicable to single data. It also plays different roles if a data set is to be used for specific data selection and if it is taken as a basis of statistical inference. Therefore, data set quality depends on the users' point of view. On the other hand, there is no identity between points of view on data set quality of the users and of database designers, the last being intended to satisfy various and divergent users' requirements. The aim ot this paper is to present, with more details, the data quality evaluation method based on vectors ordering in K-space.

---

5. Modeling the Earth's Subsurface Temperature Distribution From a Stochastic Point of View
Kirti Srivastava, National Geophysical Research Institute, India

Stochastic modeling has played an important role in the quantification of errors in various scientific investigations. In the quantification of errors one looks for the first two moments i.e. mean and variance in the system output due to errors in the input parameters. Modeling a given physical system with the available information and obtaining meaningful insight into its behavior is of vital importance in any investigation. One such investigation in Earth sciences is to understand the crustal/lithospheric evolution and temperature controlled geological processes. For this an accurate estimation of the subsurface temperature field is essential. The thermal structure of the Earth's crust is influenced by its geothermal controlling parameters such as thermal conductivity, radiogenic heat sources and initial and boundary conditions.

Modeling the subsurface temperature field is either done using a deterministic approach or the stochastic approach. In the deterministic approach the controlling parameters are assumed to be known with certainty and the subsurface temperature field is obtained. However, due to inhomogeneous and anisotropic character of the Earth's interior some amount of uncertainty in the estimation of the geothermal parameters are bound to exist. Uncertainties in these parameters may arise from the inaccuracy of measurements or lack of information available on them. Such uncertainties in parameters are incorporated in the stochastic approach and an average picture of the thermal field along with its associated error bounds is obtained.

The quantification of uncertainty in the temperatures field is obtained using both random simulation and stochastic analytical methods. The random simulation method is a numerical method in which the uncertainties in the thermal field due to uncertainties in the controlling thermal parameters are quantified. The stochastic analytical method is generally solved using the small perturbation method and closed form analytical solutions to the first two moments are obtained. The stochastic solution to the steady state heat conduction equation has been obtained for two different conditions i.e. when the heat sources are random and when the thermal conductivity is random. Closed form analytical expressions for mean and variance of the subsurface temperature distribution and the heat flow have been obtained. This study has been applied to understand the thermal state in a tectonically active region in the Indian Shield.

Track IV-B-4: Emerging Concepts of Data-Information-Knowledge Sharing

Henri Dou, Université Aix Marseille III, Marseille, France, and
Clément Paoli, Université of Marne la Vallée UMLV, Champ sur Marne, France

In various academic or professional activities the need to use distributed Data, Information and Knowledge (D-I-K) features, either as resources or in cooperative action, often becomes very critical. It is not enough to limit oneself to interfacing existing resources such as databases or management systems. In many instances, new actions and information tools must be developed. These are often critical aspects of some global changes required in existing information systems.

The complexity of situations to be dealt with implies an increasing demand for D-I-K attributes in large problems, such as environmental studies or medical systems. Hard and soft data must be joined to deal with situations where social, industrial, educational, and financial considerations are all involved. Cooperative work already calls for these intelligent knowledge management tools. Such changes will certainly induce new methodologies in management, education, and R&D.

This session will emphasize conceptual level of emerging global methodology as well as the implementation level of working tools for enabling D-I-K sharing in existing and future information systems. Issues that might be examined in greater detail include:

- Systems to develop knowledge on a cooperative basis;
- Access to D-I-K in remote teaching systems, virtual laboratories and financial aspects;
- Corporate universities (case studies will be welcomed): alternate teaching and industrial D-I-K confidentiality innovation supported by information technology in educational systems and data format interchange in SEWS (Strategic Early Warning Systems) applied to education and usage of data;
- Ethics in distance learning; and
- Cases studies on various experiments and standardization of curriculum.

1. Data Integration and Knowledge Discovery in Biomedical Databases. A Case Study

Arnold Mitnitski, Department of Medicine, Dalhousie University, Halifax, Canada
Alexander Mogilner, Montreal, Canada
Chris MacKnight, Division of Geriatric Medicine, Dalhousie University, Halifax, Canada
Kenneth Rockwood, Division of Geriatric Medicine, Dalhousie University, Halifax, Canada.

Biomedical (epidemiological) databases generally contain information about large numbers of individuals (health related variable: diseases, symptom and signs, physiological and psychological assessments, socio-economic variables etc.). Many include information about adverse outcomes (e.g.death), which makes it possible to discover links between health outcomes and other variables of interest (e.g., diseases, habits, function). Such databases also can be linked with demographic surveys that themselves contain large amounts of data aggregated by age and sex and with genetic databases. While each of the databases are usually created independently, for discrete purposes the possibility of integrating knowledge from several domains across databases is of significant scientific and practical interest. One example of linking a biomedical database (National Population Health Survey) containing more than 80,000 records of Canadian population in 1996-97 years and 38 variables (disabilities, diseases, health conditions) with mortality statistic obtained for Canadians male and female is discussed. First, the problem of the redundancy in the variables is considered. Redundancy makes it possible to derive a simple score as a generalized (macroscopic) variable that reflects both individual and group health status.

This macroscopic variable reveals a simple exponential relation with age, indicating that the process of accumulation of deficits (damage) is a leading factor causing death. The age trajectory of the statistical distribution of this variable also suggests that redundancy exhaustion is a general mechanism, reflecting different diseases. The relationship between generalized variables and the hazard (mortality) rate reveals that the latter can be expressed in terms of variables generally available from any cross-sectional database. In practical terms, this means that the risk of mortality might readily be assessed from standard biomedical appraisals collected on other grounds. This finding is an example of how knowledge from different data sources can be integrated to common good ends. Additionally, Internet related technologies might provide ready means to facilitate interoperability and data integration.

---

2. A Framework for Semantic Context Representation of Multimedia Resources

Weihong Huang , Yannick Prié , Pierre-Antoine Champin, Alain Mille
LISI, Université Claude Bernard Lyon 1, France

With the explosion of online multimedia resources, requirement of intelligent content-based multimedia service increases rapidly. One of the key challenges in this area is semantic contextual knowledge representation of multimedia resources. Although current image and video indexing techniques enable efficient feature-based operation on multimedia resources, there still exists a "semantic gap"between users and the computer systems, which refers to the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data has for a user in a given situation.

In this paper, we present a novel model: annotation graph (AG) for modeling and representing contextual knowledge of various types of resources such as text, image, and audio-visual resources. Based on the AG model, we attempt to build an annotation graph framework towards bridging the "semantic gap" by offering universal flexible knowledge creation, organization and retrieval services to users. In this framework, users will not only benefit from semantic query and navigation services, but also be able to contribute in knowledge creation via semantic annotation.

In the AG model, four types of concrete description elements are designed for concrete descriptions in specific situations, while two types of abstract description elements are designed for knowledge reusing in different situations. With these elements and directed arcs between them, contextual knowledge at different semantic levels could be represented through semantic annotation. Within the global annotation graph constructed by all AGs, we provide flexible semantic navigation using derivative graphs (DG) and AG. DGs enable complement contextual knowledge representation to AGs by focusing on different types of description elements. Towards semantic query, we present a potential graph (PG) tool to help users visualize query requests as PGs, and execute queries by performing sub-graph matching with PGs. Prototype system design and implementation aim at an integrated user-centered semantic contextual knowledge creation, organization and retrieval system.

3. Passer de la représentation du présent à la vision prospective du futur - Du Technology Forecast au Technology Foresight
Henri Dou, CRRM, Université Aix Marseille III, Centre Scientifique de Saint Jérôme, France
Jin Zhouiyng, Institute of Techno-Economics, Chinese Academy of Social Science (CASS), China

De nos jours, le passage du système technology forecast au système technology foresight est inévitable pour éviter que le développement scientifique ne soit qu'orienté verticalement au détriment des retombées possibles (positives ou négatives) au niveau de la Société. Dans cette communication les auteurs aborderont les aspects méthodologiques de cet passage ainsi que les différentes étapes qui ont jalonnées depuis 1930 cette évolution. Les analyses réalisées par différents payes seront présentées, avec un panorama international des actions en cours dans ce domaine.

Le concept Technology Foresight sera ensuite introduit dans la méthodologie de l'Intelligence Compétitive Technique ou Economique afin de créer pour des entreprises une vision du développement soutenable et éthique pour créer de nouveaux avantages.

La mise en œuvre internationale du concept, tant au plan Européen (6ième PCRD), qu'au niveau de la déclaration de Bologne (Juin 1999), et des actions menées au Japon ou en Chine (China 2020) sera analysée.

4. Mise en place d'un système dynamique et interactif de gestion d'activité et de connaissances d'un laboratoire
Mylène Leitzelman : Intelligence Process SAS, France
Valérie Léveillé : Case 422 Centre Scientifique de Saint-Jérôme, France
Jacky Kister : UMR 6171 S.C.C - Faculté des Sciences et Techniques de St Jérôme, France

Il s'agit de mettre en place de façon expérimentale et pour le compte de l'UMR 6171 associé au CRRM, un système interconnecté de gestion d'activité et de connaissances pour gérer l'activité scientifique d'une unité de recherche. Ce système sera doté de modules de visualisation synthétique, statistiques et cartographiques s'appuyant sur des méthodologies de datamining et de bibliométrie. Le point clé de ce système sera de proposer en même temps un outil de gestion stratégique et d'organisation d'un laboratoire et un outil permettant la compilation interlaboratoires pour en faire un outil d'analyse ou de stratégie à une plus grande échelle en laissant des accès plus ou moins libres pour que des agents extérieurs puissent à partir des données générer des indicateurs de performance, de valorisation, de qualité des productions scientifiques et de relations laboratoire/entreprises.

5. La dimension éthique de la relation pédagogique dans la formation à distance
M. Lebreton, C. Riffaut, H. Dou, Faculté des sciences et techniques de Marseille Saint-Jérôme (CRRM), France

De tous temps, enseigner a signifié être mis en relation avec quelqu'un dans le but de lui apprendre quelque chose. Le lien qui va unir le formateur à l'apprenant sera le savoir. Se forme ainsi un triangle éducatif[1] dont les branches constitutent la(les) relation(s) pédagogique(s).
Pour pouvoir activer cette structure, il est nécessaire que chaque acteur connaisse avec clarté et précision ses propres motivations et ses objectifs. Par ailleurs, il paraît évident que pour transmettre et acquérir des savoirs, il est nécessaire que les partenaires du processus d'apprentissage partagent un certain nombre de valeurs communes, véritable ciment de l'acte éducatif.

Au triangle sus-mentionné, correspond un triangle éthique où à chaque sommet on peut placer l'intitulé des missions éducatives : instruire, sociabiliser et qualifier.

Instruire, c'est avant tout acquérir des connaissances. Sociabiliser, c'est surtout acquérir des valeurs. Qualifier, c'est intégrer dans une organisation productive.

Ces deux triangles ont fonctionné pendant des siécles et l'arrivée des nouvelles technologies multimédia et de la communication a destructuré la règle des trois unités -le temps, le lieu et l'action[2]. Cet ensemble est en train de se fissurer pour donner naissance à un nouveau paysage scolaire où la classe ne sera plus le seul lieu de formation, où le transfert des savoirs pourra être fait à tout moment et en tout lieu et où enfin l'action pédagogique sera individualisée et individualisable.

Dans ce nouveau contexte, la relation pédagogique dans la formation à distance va nécessiter la mise en œuvre de nouvelles compétences techniques, intellectuelles et sociales ou éthiques.
Pour pouvoir aborder ces nouveaux défis, il paraît nécessaire de chercher à savoir dans un premier temps en quoi l'éthique peut nous aider à comprendre de quelles manières ont évolué les dispositifs fondamentaux de production des savoirs et les changements intervenus dans le système de transfert des connaissances tout en se préoccupant de l'adaptation et de la nécessaire réactualisation pemanente des contenus éducatifs qui s'imposeront dorénavant.

Par la suite, le questionnement éthique doit conduire à aborder les conséquences liées à la dépersonnalisation de la relation d'apprentissage. A cet effet, il semble opportun de chercher à répondre à deux questions fondamentales. L'une a trait au formateur, est-il encore maître du processus de sociabilisation et s'interroger par la suite pour savoir si la formation à distance a encore des valeurs et dans ce cas quelles sont-elles?.

L'autre va concerner l'apprenant d'une part et l'on va s'interesser à ce qu'il advient de son identité dans l'univers du numérique et du virtuel et d'autre part chercher à savoir quel est son salut face à la marchandisation des connaissances et à l'accaparement des savoirs par des empires informationnels.

L'ensemble de ces interrogations éthiques peut permettre de commencer à trouver des débuts de solutions à des problématiques sans frontière et d'une complexité redoutable où cohabitent désormais le rationnel et l'iirtionnel, le matériel et l'immatériel, le personnel et l'impersonnel le tout immergé dans le numérique, fondement de la virtualité.

1. Le triangle pédagogique, J. Houssaye, Berne, Ed. Peter Lang
2. Rapport au Premier ministre du Sénateur A. Gérard, 1997

Data Policy

Track I-D-4: The Public Domain in Scientific and Technical Data: A Review of Recent Initiatives and Emerging Issues

Chair: Paul F. Uhlir, The National Academies, USA

The body of scientific and technical data and other information in the public domain is massive and has contributed broadly to the scientific, economic, social, cultural, and intellectual vibrancy of the entire world. The "public domain" may be defined in legal terms as sources and types of data and information whose uses are not restricted by statutory intellectual property regimes and that are accordingly available to the public without authorization. In recent years, however, there have been growing legal, economic, and technological pressures on public-domain information-scientific and otherwise-forcing a reevaluation of the role and value of the public domain. Despite these pressures, some well-established mechanisms for preserving the public domain in scientific data exist in the government, university, and not-for-profit sectors. In addition, very innovative models for promoting various public-domain digital information resources are now being developed by different groups in the scientific, library, and legal communities. This session will review some of the recent initiatives for preserving and promoting the public domain in scientific data within CODATA and ICSU, the US National Academies, OECD, UNESCO, and other organizations, and will highlight some of the most important emerging issues in this context.

---

1. International Access to Data and Information
Ferris Webster, University of Delaware, USA

Access to data and information for research and education is the principal concern of the ICSU/CODATA ad hoc Group on Data and Information. The Group tracks developments by intergovernmental organizations with influence over data property rights. Where possible, the Group works to assure that the policies of these organizations recognize the public good to be derived by assuring access to data and information for research and education.

A number of international organizations have merited attention recently. New proprietary data rights threaten to close off access to data and information that could be vital for progress in research. The European Community has been carrying out a review of its Database Directive. The World Meteorological Organization's resolution on international exchange of meteorological data has been the subject of continuing debate. The Intergovernmental Oceanographic Commission is drafting a new data policy that may have constraints that are parallel to those of the WMO. The World Intellectual Property Organization has had a potential treaty on databases simmering for several years.

The latest developments in these organizations will be reviewed, along with the activities of the ICSU/CODATA Group.

2. The OECD Follow up Group on Issues of Access to Publicly Funded Research Data: A Summary of the Interim Report
Peter Arzberger, University of California at San Diego, USA

This talk will present a summary of the interim report of the OECD Follow up Group on Issues of Access to Publicly Funded Research Data. The Group's efforts have origins in the 3rd Global Research Village conference in Amsterdam, December 2000. In particular, it will include issues of global sharing of research data. The Group has conducted case studies of practices across different communities, and looked at factors such as sociological, economic, technological and legal issues that either enhance or inhibit data sharing. The presentation will also address issues such as data ownership and rights of disposal, multiple uses of data, the use of ICT for widening the scale and scope of data-sharing, effects of data-sharing on the research process, and co-ordination in data management. The ultimate goal of the Group is to articulate principles, based on best practices that can be interpreted into the science policy arena. Some initial principles will be discussed. Questions such as the following will be addressed:

- What principles should govern science policy in this area?
- What is the perspective of social informatics in this field?
- What role does the scientific community play in this?

It is intended that this presentation will generate discussion and feedback on key points of the Group's interim report.

3. An Overview of Draft UNESCO Policy Guidelines for the Development and Promotion of Public-Domain Information
John B. Rose, UNESCO, Paris, FRANCE
Paul F. Uhlir, The National Academies, Washington, DC, USA

A significantly underappreciated, but essential, element of the information revolution and emerging knowledge society is the vast amount of information in the public domain. Whereas the focus of most policy analyses and law making is almost exclusively on the enhanced protection of private, proprietary information, the role of public-domain information, especially of information produced by the public sector, is seldom addressed and generally poorly understood.

The purpose of UNESCO's Policy Guidelines for the Development and Promotion of Public-Domain Information, therefore, is to help develop and promote information in the public domain at the national level, with particular attention to information in digital form. These Policy Guidelines are intended to better define public-domain information and to describe its role and importance, specifically in the context of developing countries; to suggest principles that can help guide the development of policy, infrastructure and services for provision of government information to the public; to assist in fostering the production, archiving and dissemination of an electronic public domain of information for development, with emphasis on ensuring multicultural, multilingual content; and to help promote access of all citizens, especially including disadvantaged communities, to information required for individual and social development. This presentation will review the main elements of the draft Policy Guidelines, with particular focus on scientific data and information in the public domain.

Complementary to, but distinct from, the public domain are the wider range of information and data which could be made available by rights holders under specific "open access" conditions, as in the case of open source software, and the free availability of protected information for certain specific purposes, such as education and science under limitations and exceptions to copyright (e.g., "fair use" in U.S. law). UNESCO is working to promote international consensus on the role of these facilities in the digital age, notably through a recommendation under development on the "Promotion and Use of Multilingualism and Universal Access to Cyberspace," which is intended to be presented to the World Summit on the Information Society to be organized in Geneva (2003) and Tunis (2005), as well as a number of other relevant programme actions which will also be presented at the Summit.

4. Emerging Models for Maintaining the Public Commons in Scientific Data
Harlan Onsrud, University of Maine, USA

Scientists need full and open disclosure and the ability to critique in detail the methods, data, and results of their peers. Yet scientific publications and data sets are burdened increasingly by access restrictions imposed by legislative acts and case law that are detrimental to the advancement of science. As a result, scientists and legal scholars are exploring combined technological and legal workarounds that will allow scientists to continue to adhere to the mores of science without being declared as lawbreakers. This presentation reviews three separate models that might be used for preserving and expanding the public domain in scientific data. Explored are the technological and legal underpinnings of Research Index, the Creative Commons Project and the Public Commons for Geographic Data Project. The first project relies heavily on protections granted to web crawlers under the U.S. Digital Millennium Copyright Act while the latter two rely on legal approaches utilizing open access licenses.

5. Progress, Challenges, and Opportunities for Public-Domain S&T Data Policy Reform in China
Liu Chuang, Chinese Academy of Sciences, Beijing, China

China has experienced four different stages for public-domain S&T data management and policy during the last quarter century. Before 1980, most of the government funded S&T data were free to be accessed, and the services received a good reputation from the scientific community. Most of these data were recorded on paper media, however, and took time to be accessed.

With the computer developments in the early 1980s, digital data and databases increased rapidly. The data producers and holders began to realize that the digital data could be an important resources for the scientific activities. The policy to charge fees for data access gained prominence between the early 1980s and approximately 1993. During this time period, China experienced new problems in S&T data management. For example, there was an increase of parallel work in database development and in data controlled by individual persons with a high risk of losing the data, and the price of access to data became very expensive in most cases.

In the 1994-2000 period, members of the scientific community asked for data policy reform, and for lower costs of access to government funded databases for non-profit applications. The Ministry of Science and Technology (MOST) set up a group to investigate China's S&T data sharing policies and practices.

A new program for S&T data sharing was initiated by MOST in 2001. This was a major milestone for enhancing access to and the application of public-domain S&T data. This new program, along with the current development of a new data access policy and support system, is expected to be greatly expanded during next decade.

Track IV-A-4: Confidentiality Preservation Techniques in the Behavioral, Medical and Social Sciences

D. Johnson, Building Engineering and Science Talent, San Diego, CA, USA
J. McArdle, Dept. of Physchology, University of Virginia, USA
Julie Kaneshiro, National Institutes of Health, USA
Kurt Pawlik, Psychologisches Institut I, Universität Hamburg, Germany
Michel Sabourin, Université de Montréal, Canada

In the behavioral and social sciences and in medicine, the movement to place data in electronic databases is hampered by considerations of confidentiality. The data collected on individuals by scientists in these areas of research are often highly personal. In fact, it is often necessary to guarantee potential research participants that the data collected on them will be held in strictest confidence and that their privacy will be protected. There has even been debate in these sciences about whether data collected under a formal confidentiality agreement can be placed in a database, because such use might constitute a use of the data to which the research participants did not consent.

The members of this panel will discuss a broad range of techniques that are being used across the behavioral and social sciences and medicine to protect the confidentiality of individuals whose data are entered into an electronically accessible database. Among the highly controversial data to which these techniques are being applied are data on accident avoidance by pilots of commercial aircraft and data on medical errors. The stakes in finding ways to use these data without violating confidentiality are high, since the payoff from learning how to reduce airplane accidents and medical mistakes is saved lives.

Standard techniques for separating identifier information from data, as well as less common techniques such as the introduction of systematic error in data, will be discussed. Despite the methods that are in place and those that are being experimented with, there is evidence that even sophisticated protection techniques may not be enough. The group will conclude its session with a discussion of this challenge.

1. Issues in Accessing and Sharing Confidential Survey and Social Science Data
Virginia A. de Wolf, USA

Researchers collect data under pledges of confidentiality. The US federal statistical system has established practices and procedures that enable others to access data it collects. The two main methods the federal statistical agencies use are to restrict the content of the data (termed "restricted data") and to restrict the conditions under which the data can be accessed, i.e., at what locations, for what purposes (term "restricted access").

Data sharing practices in the various social science disciplines vary. For example, codes of ethics of some social science disciplines encourage sharing (e.g., the American Association for Public Opinion Research) while others do not. In the US both of the institutions that fund the bulk of the social science research, the National Institutes of Health and the National Science Foundation, have statements on data sharing.

This presentation will review the practices, procedures, issues, etc., of US federal statistical agencies in allowing access to the data they collect. It will highlight the activities of US federal interagency committees and will conclude with a discussion of the applicability of the experience of the US federal statistical system to the academic social science community.

2. Contemporary Statistical Techniques for Closing the "Confidentiality Gap" in Behavioral Science Research
John J. McArdle, University of Virginia, USA
David Johnson, Building Engineering and Science Talent, San Diego, CA, USA

Over the past three decades, behavioral scientists have become acutely aware of the need for both the privacy of research participants and the confidentiality of research data. During this same time period, knowledgeable researchers have created a variety of methods and procedures to insure confidentiality. But many of the best techniques used were not designed to permit the sharing of research data with other researchers outside of the initial data collection group. Since a great deal of behavioral science data collected at the individual level require such protections they cannot easily be shared with others in a confidential way. These practical problems have created a great deal of confusion and a kind of "confidentiality gap" among researchers and participants alike. This presentation will review some available "statistical" approaches to deal with these problems, and examples will be drawn from research projects on human cognitive abilities. These statistical techniques range from the classical use of replacement or shuffled records to more contemporary techniques based on multiple imputations. In addition, new indices will be used to relate the potential loss of data accuracy versus the loss of confidentiality. These indices will help researchers define the confidentiality gap in their own and any other research project.

References

1. Feinberg, S.E, & Willenborg, L. C.R.J. (1998). Special issue on "Disclosure limitation methods for protecting confidentiality of statistical data." Journal of Official Statistics, 14 (4), 337-566.
2. Willenborg, L. C.R.J. & de Waal, T. (2001). Elements of statistical disclosure control. Lecture Notes in Statistics, 155. New York: Springer-Verlag.
3. Clubb, J.M., Austin, E.W., Geda, C.L. & Traugott, M.W. (1992). Sharing research data in the social sciences. In G. H. Elder, Jr., E. K. Pavalko & E. C. Clipp. Working with Archival Data: Studying Lives (pp. 39-75). SAGE Publications.
4. Willenborg, L. C.R.J. & de Waal, T. (1996). Statistical disclosure control in practice. Lecture Notes in Statistics, 111. New York: Springer-Verlag.

3. NASA Aviation Safety Reporting System (ASRS)
Linda J. Connell, NASA Ames Research Center, USA

In 1974, the United States experienced a tragic aviation accident involving a B-727 on approach to Dulles Airport in Virginia. All passengers and crew were killed. The accident was classified as a Controlled Flight Into Terrain event. During the NTSB accident investigation, it was discovered from ATC and cockpit voice recorder tapes that the crew had become confused over information regarding the approach instructions, both in information provided in approach charts and the ATC instruction "cleared for the approach". It was discovered that another airline had experienced a similar chain of events, but they detected the error and increased their altitude. This action allowed them to miss the on-coming mountain. The second event would be classified as an incident. The benefit of the information spread rapidly in this airline, but had not reached other airlines. As a result of the NTSB findings, the FAA and NASA created the Aviation Safety Reporting System in 1976. The presentation will describe the background and principles that guide the operation of the ASRS. The presentation will also include descriptions of the uses of and products from approximately 490,000 incident reports.

Technical Demonstrations

Track II-D-2: Technical Demonstrations

Chairs: Richard Chinman, University Corporation for Atmospheric Research, Boulder, CO, USA and
Robert S. Chen, CIESIN, Columbia University, USA

---

1. World Wide Web Mirroring Technology of the World Data Center System
David M. Clark, World Data Center Panel, NOAA/NESDIS, USA

The widespread implementation and acceptance of the World Wide Web (WWW) has changed many facets of the techniques by which Earth and environmental data are accessed, compiled, archived, analyzed and exchanged. The ICSU World Data Centers, established over 50 years ago, are beginning to use this technology as they evolve into a new way of operations. One key element of this new technology is known as WWW "mirroring." Strictly speaking, mirroring is reproducing exactly the web content from one site to another at physically separated location. However, there are other types of "mirroring" which uses the same technology, but are different in appearance and/or content of the site. The WDCs are beginning to use these three types of mirroring technology to encourage new partners in the WDC system. These new WDC partners bring a regional diversity or a discipline specific enhancement to the WDC system. Currently there are ten sites on five continents mirroring a variety of data types using the different modes of mirroring technology. These include paleoclimate data mirrored in the US, Kenya, Argentina and France, and space environment data mirrored in the US, Japan, South Africa, Australia and Russia. These mirror sites have greatly enhanced the exchange and integrity of the respective discipline databases. A demonstration of this technology will be presented.

---

2. Natural Language Knowledge Discovery: Cluster Grouping Optimization
Robert J. Watts, U.S. Army Tank-automotive and Armaments Command, National Automotive Center, USA
Alan L. Porter, Search Technology, Inc. and Georgia Tech, USA
Donghua Zhu, Beijing Institute of Technology, China

The Technology Opportunities Analysis of Scientific Information System (Tech OASIS), commercially available under the trade name VantagePoint, automates the identification and visualization of relationships inherent in sets (i.e., hundreds or thousands) of literature abstracts. A Tech OASIS proprietary approach applies principal components analysis (PCA), multi-dimensional scaling (MDS) and a path-erasing algorithm to elicit and display clusters of related concepts. However, cluster groupings and visual representations are not singular for the same set of literature abstracts (i.e., user selection of the items to be clustered and the number of factors to be considered will generate alternative cluster solutions and relationships displays). Our current research, the results of which shall be demonstrated, seeks to identify and automate selection of a "best" cluster analysis solution for a set of literature abstracts. How then can a "best" solution be identified? Research on quality measures of factor/cluster groups indicates that those that appear promising are entropy, F measure and cohesiveness. Our developed approach strives to minimize the entropy and F measures and maximize cohesiveness, and also considers set coverage. We apply this to automatically map conceptual (term) relationships for 1202 abstracts concerning "natural language knowledge discovery."

3. ADRES: An online reporting system for veterinary hospitals
P.K. Sidhu and N.K. Dhand, Punjab Agricultural University, India

An animal husbandry department reporting system (ADRES) has been developed for online submission of monthly progress reports of veterinary hospitals. It is a database prepared under Microsoft Access 2000, which has records of all the veterinary hospitals and dispensaries of animal husbandry department, Punjab, India. Every institution has been given a separate ID. The codes for various infectious diseases have been selected according to the codes given by OIE (Office International des Epizooties). In addition to reports about disease occurrence, information can also be recorded for progress of insemination program, animals slaughtered in abattoirs, animals exported to other states and countries, animal welfare camps held and farmer training camps organized etc. Records can be easily compiled on sub-division, district and state basis and reports can be prepared online for submission to Government of India. It is visualized that the system may make the reports submission digital, efficient and accurate. Although, the database has been primarily developed for Punjab State, other states of India and other countries may also easily use it.

4. PAU_Epi~AID: A relational database for epidemiological, clinical and laboratory data management
N.K. Dhand, Punjab Agricultural University, India

A veterinary database (Punjab Agricultural University Epidemiological Animal disease Investigation Database, PAU_ Epi~AID) has been developed to meet the requirements of data management during outbreak investigations, monitoring and surveillance, clinical and laboratory investigations. It is based on Microsoft Access 2000 and includes a databank of digitalized information of all states and union territories of India. Information of districts, sub divisions, veterinary institutions and important villages of Punjab (India) has also been incorporated, every unit being represented by an independent numeric code. More than 60 interrelated tables have been prepared for registering information on animal disease outbreaks, farm data viz. housing, feeding, management, past disease history, vaccination history etc. and animal general information, production, reproduction and disease data. Findings of various laboratories such as bacteriology, virology, pathology, parasitology, molecular biology, toxicology, serology etc. can also be documented. Data can be easily entered in simple forms hyper-linked to one another, which allow queries and reports preparation at click of mouse. Flexibility has been provided for additional requirements due to diverse needs. The database may be of immense use in data storage, retrieval and management in epidemiological institutions and veterinary clinics.

5. Archiving Technology for Natural Resources and Environmental Data in Developing Countries, A Case Study in China
Wang Zhengxing, Chen Wenbo, Liu Chuang, Ding Xiaoqiang, Chinese Academy of Sciences, China

Data archiving has long been regarded as a less important sector in China. As a result, there is no long-term commitment at the national level to preserve natural resources data, and usually smaller budgets for data management than for research. Therefore, it is essential to develop a feasible strategy and technology to manage the exponential growth of the data. The strategy and technology should be cost-saving, robust, user-friendly, and sustainable in the long run. A PC-based system has been developed to manage satellite imagery, Geographic Information System (GIS) maps, tabular attribute data, and text data. The data in text format include data policies compiled from international, national, and regional organizations. Full documentation on these data are on-line and free to download. Only metadata and documentation are on-line for GIS maps and tabular data; the full datasets are distributed by CD-ROM, e-mail, or ftp.

Remote sensing data are often too expensive for developing countries. An agreement has been reached between GCIRC and remote sensing receiving station vendors. According to the agreement, GCIRC can freely use the remote sensing data (MODIS) from the receiving station, conditional on making their system available to demonstrate to potential buyers. This assures the most important data source for archiving. Considering the huge

volumes of data and limited PC capacity, only quick-look images and metedata are permanently on-line. Users can search for data by date, geolocation, or granule. Full 1B images are updated daily and kept on-line for one week; users can download the recent data for free. All raw data (direct broadcast) and 1B images are archived on CD-ROMs, which are easy to read using a personal computer.

---

6. Delivering interdisciplinary spatial data online: The Ramsar Wetland Data Gateway
Greg Yetman and Robert S. Chen, Columbia University, USA

Natural resource managers and researchers around the world are facing a range of cross-disciplinary issues involving global and regional environmental change, threats to biodiversity and long-term sustainability, and increasing human pressures on the environment. They must increasingly harness a range of socioeconomic and environental data to better understand and manage natural resources at local, regional, and global scales.

This demonstration will illustrate an online information resource designed to help meet the interdisciplinary data needs of scientists and resource managers concerned with wetlands of international importance. The Ramsar Wetland Data Gateway, developed in collaboration with the Ramsar Bureau and Wetlands International, combines relational database technology with interactive mapping tools to provide powerful search and visualization capabilities across a range of data from different sources and disciplines. The Gateway is also being developed to support interoperable data access across distributed spatial data servers.

Large Data Projects

Track I-D-3: Land Remote Sensing - Landsat Today and Tomorrow

Chairs: Hedy Rossmeissl and John Faundeen, US Geological Survey, USA

Scientists in earth science research and applications and map-makers have for many years been avid users of remotely sensed Landsat data. The use of remote sensing technology, and Landsat data in particular, is extremely useful for illustrating: current conditions and temporal change for monitoring and assessing the impacts of natural disasters; aiding in the management of water, biological, energy, and mineral resources; evaluating environmental conditions; and enhancing the quality of life for citizens across the globe. The size of the image files, however, raises a variety of data management challenges. This session will focus specifically on the 30-year experience with Landsat image data and will examine four components: 1) image tasking, access, and dissemination, 2) applications and use of the imagery, 3) data archiving, and 4) the future of the Landsat program.

---

1. Tasking, Archiving & Dissemination of Landsat Data
Thomas J. Feehan, Canada Centre for Remote Sensing, Natural Resources Canada, Canada

The Canada Centre for Remote Sensing of Natural Resources Canada (CCRS) operates two satellite ground receiving stations, the Prince Albert Satellite Station located in Prince Albert Saskatchewan and the Gatineau Satellite Station located in Cantley, Quebec. The CCRS stations provide a North American data reception capability, acquiring data to generate knowledge and information critical to resource use decision making on local, regional, national and global scales. CCRS' primary role is to provide data related to land resources and climate change, contributing to sustainable land management in Canada.

Operating in a multi-mission environment, including LANDSAT, the CCRS stations have accumulated an archive in excess of 300 TeraBytes, dating back to 1972, when CCRS started receiving LANDSAT-1 (ERST-1) data at the Prince Albert Satellite Station. Data are made available to support near-real time applications including ice monitoring, forest fire monitoring and mapping, as well as non-real time applications such as climate change, land use and topographic mapping. LANDSAT MSS, TM and ETM+ data constitute a significant portion of the CCRS archive holdings.

In addition to Canadian Public Good data use, a spin-off benefit includes the commercial exploitation by a CCRS distributor and value-added services network.

---

2. The Work of the U.S. National Satellite Land Remote Sensing Data Archive Committee: 1998 - 2000
Joanne Irene Gabrynowicz, National Remote Sensing and Space Law Center, University of Mississippi School of Law, USA

Earth observation data have been acquired and stored since the early 1970s. One of the world's largest, and most important, repositories for land satellite data is the Earth Resources Observation Systems (EROS) Data Center (EDC). It is a data management, systems development, and research field center for the U.S. Geological Survey's (USGS) National Mapping Discipline in Sioux Falls, South Dakota, USA. It was established in the early 1970s and in 1992, the U.S. Congress established the National Satellite Land Remote Sensing Data Archive at EDC. Although data have been acquired and stored for decades, the world's remote sensing community has only recently begun to address long-term data preservation and access. One such effort was made recently by remote sensing leaders from academia, industry and government as members of a federal advisory committee from 1998 to 2000. This presentation provides a brief account of the Committee's work product.

3. An Overview of the Landsat Data Continuity Mission (LDCM)
Bruce K. Quirk and Darla M. Duval*, U.S. Geological Survey EROS Data Center, USA
Jeffrey G. Masek, Douglas McCuistion, James Irons, NASA Goddard Space Flight Center, USA

Since 1972 the Landsat program has provided continuous observations of the Earth's land areas, giving researchers and policy makers an unprecedented vantage point for assessing global environmental changes. The analysis of this record has driven a revolution in terrestrial remote sensing over the past 30 years. Landsat 7 was successfully launched in 1999 and returned operation of the Landsat program to the U.S. government. This presentation describes plans for the follow-on to Landsat 7, the Landsat Data Continuity Mission (LDCM), which has a planned launch date of late 2005. The scientific need for Landsat-type observations has not diminished through time. Changes in global land cover have profound implications for the global carbon cycle, climate, and functioning of ecosystems. Furthermore, these changes must be monitored continually in order to link them to natural and socioeconomic drivers. Landsat observations play a key role, because they occupy that unique part of the spatial-temporal domain that allows human-induced changes to be separated from natural changes. Coarse resolution sensors such as MODIS and AVHRR are ideal for monitoring the daily and weekly changes in global biophysical conditions, but lack the resolution to accurately measure the amount and origin of land cover change. High-resolution commercial systems, while valuable for validation, cannot acquire sufficient global data to meet scientific monitoring needs. Landsat-type observations fill this unique niche. A joint effort between the National Aeronautics and Space Administration (NASA) and the U.S. Geological Survey (USGS), LDCM will continue the Landsat legacy by incorporating enhancements that reduce system cost and improve data quality. Following the 1992 Land Remote Sensing Policy Act, the LDCM seeks a commercially owned and operated system selected through a competitive procurement. Unlike earlier Landsat commercialization efforts, however, the LDCM procurement is based on a rigorous Science Data Specification and Data Policy, which seek to guarantee the quantity and quality of the data, while preserving reasonable cost and unrestricted data rights for end users. Thus the LDCM represents a unique opportunity for NASA and USGS to provide science data in partnership with private industry, to reduce cost and risk to both parties, while concurrently creating an environment to expand the commercial remote sensing market. The Data Specification requires provision of 250 scenes per day, globally distributed, with modest improvements in radiometric signal-to-noise (SNR) and dynamic range. Two additional bands may be included following trade studies during 2002: an "ultra-blue" band centered at 443 nm for coastal and aerosol studies, and a band at 1375 nm for cirrus cloud detection. Two thermal bands with a resolution of 120 meters may also be included. In addition to the rationale for science data continuity, this presentation will give additional details on the LDCM specification, mission concept, and status.

* Raytheon. Work performed under U.S. Geological Survey contract 1434-CR-97-CN-40274.

4. Current Applications of Landsat 7 Data in Texas
Gordon L. Wells, Center for Space Research, The University of Texas at Austin, USA

The rapid delivery of timely information useful to decision makers is one of the primary goals of the data production and application programs developed by the Mid-American Geospatial Information Center (MAGIC) located at the University of Texas at Austin's Center for Space Research. In a state the size and nature of Texas, geospatial information collected by remote sensing satellites can assist a broad range of operational activities within federal, state, regional and local government departments. In the field of emergency management, the state refreshes its imagery basemap using Landsat 7 data on a seasonal basis to capture the locations of recent additions to street and road networks and new structures that might be vulnerable to wildfires or flashfloods. Accurately geolocated satellite imagery can be incorporated into the geographic information system used by the Governor's Division of Emergency Management much more rapidly than updated records received from the department of transportation or local entities. For many activities involving the protection and enhancement of natural resources, Landsat 7 data offer the most economic and effective means to address problems that affect large areas. Invasive species detection and eradication is a current concern of the Texas Department of Agriculture, Texas Soil and Water Conservation Board and the Upper Colorado River Authority. Invasive saltcedar is one noxious species that can be identified and

removed with the help of satellite remote sensing. The information required by policy makers may extend beyond state borders into regions where satellite reconnaissance is the only practical tool available. For international negotiations involving the shared water resources of Texas and Mexico, satellite imagery has made a valuable contribution to the monitoring of irrigation activities and the local effects of drought conditions. In the future, there will be increasing concentration on shortening the time lag between the collection of instrument data by MAGIC's satellite receiving station and final product delivery in the projection, datum and file format required for immediate inclusion into operational analyses by the various agencies in the region.

---

5. Development of Land Cover Database of East Asia

Wang Zhengxing, Zhao Bingru, Liu Chuang, Global Change Information and Research Center, Institute of Geography and Natural Resource Research, Chinese Academy of Sciences, China

Land cover plays a major role in a wide range of fields from global change to regional sustainable development. Although land cover has dramatically changed over the last few centuries, util now there has been no consistent way of quantifying the changes globally (Nemani, and Running, 1995). Land cover dataset currently used for parameterization of global climate models are typically derived from a range of preexisting maps and atlases (Olson and Watts, 1982; Matthews, 1983; Wilson and Henderson-Sellers, 1985), this approach has several limitations (A. Strahler and J. Townshend, 1996). Another important data source is statistical report, but some statistical land cover data seems unreliable. At present, the only practical way to develop land cover dataset consistently, continuously, and at globally is satellite remote sensing. This is also true for the development of land cover dataset of East Asia.

The 17-class IGBP land cover unit includes eleven classes of natural vegetation, three classes of developed and mosaic lands, and three classes of non-vegetated lands. This system may be useful at global level, but there is a very serious shortcoming: only one class for arable land. Since the arable land is the most dynamic and important area of the man-nature system, it is essential to characterize arable land sub-system to more details.

There are still some potentials for finer classification in current 1-km AVHRR-NDVI data sets. A decision tree classifier is used to transfer all input data into various pre-defined classes. The key to accurate interpretation is to identify more reliable links (decision rules) between input data and output classes. The basic theory under the decision tree is that any land cover class should be an identical point determined by a multi-dimensional spaces, including multi temporal NDVI, phenology, ecological region, DEM, census data etc. The preliminary research shows that stratification using ecological region and DEM can simplify the decision tree structure and yield more meaningful classes in China's major agricultural regions. Arable land cover may be classified at two levels, first level describes how many times the crops are planted, and second level the crop characteristics.

The current land cover classification based on 1-km AVHRR-NDVI data sets still have serious limitations for parameterization of some models. The nominal 1-km spatial resolution images may produce quite a lot mixed pixels, but some models need pure pixel, e.g. DNDC model. However, the coming 250-m MODIS-EVI data set will narrow the gap between model need and data supply to some extent. Using the approaches developed from AVHRR, MODIS will yield more reliable land cover data of East Asia.

Roundtable

Track II-D-1: Roundtable Discussion on Preservation and Archiving of Scientific and Technical Data in Developing Countries

Chair: William Anderson, Praxis101, Rye, NY, USA

Session Organizers: William Anderson, US National Committee for CODATA
Steve Rossouw, South African National Committee for CODATA
Liu Chuang, Chinese Academy of Sciences, Beijing, China
Paul F. Uhlir, US National Committee for CODATA

A Working Group on Scientific Data Archiving was formed following the 2000 CODATA Conference. The primary objective of this Working Group has been to create a focus within CODATA on the issues of scientific and technical data preservation and access. This Working Group, co-chaired by William Anderson and Steve Rossouw, has co-organized a workshop on data archiving with the South African National Research Foundation in Pretoria in May 2002. The Working Group is preparing a report of its activities from 2001-2002.

Another initiative of the Working Group has been to propose the creation of a CODATA "Task Group on Preservation and Archiving of S&T Data in Developing Countries." The proposed objectives of that Task Group are to: promote a deeper understanding of the conditions in developing countries with regard to long-term preservation, archiving, and access to scientific and technical (S&T) data; advance the development and adoption of improved archiving procedures, technologies, standards, and policies; provide an interdisciplinary forum and mechanisms for exchanging information about S&T data archiving requirements and activities, with particular focus on the concerns of developing countries; and publish and disseminate broadly the results of these efforts. The proposed Task Group would be co-chaired by William Anderson and Liu Chuang.

An additional related proposal of the Working Group has been to create a Web portal on archiving and preservation of scientific and technical data and information. This portal, which would be developed jointly by CODATA with the International Council for Scientific and Technical Information and other interested organizations, would provide information about and links to online:

- Scientific and technical data and information archiving procedures, technologies, standards, and policies;
- Discipline-specific and cross-disciplinary archiving projects and activities; and
- Expert points of contact in all countries, with particular attention to those in developing countries.

Reports on all these activities will be given at the Roundtable and will then be discussed with the individuals who attend this session.

Overview and Grand Challenges

Thursday, 3 October 1200 – 1300

Chair: Fedor Kuznetzov, Institute of Inorganic Chemistry, Novosibirsk, Russia

Preserving Scientific Data: Supporting Discovery into the Future
John Rumble, CODATA President

A wide variety of methods have been used to save and preserve scientific data for thousands of years. The physical nature of these means and the inherent difficulties of sharing the physical media with others who need the data ha e been major barriers in advancing research and scientific discovery. The information revolution is changing this in many significant ways; ease of availability, breadth of distribution, size and completeness of data sets, and documentation. As a consequence, scientific discovery itself is changing now, and in the future, perhaps even more dramatically. In this talk I will review some historical aspects of data preservation and the use of data in discovery. And I will provide some speculations on how preserving data digitally might revolutionize scientific discovery.

Poster Session Abstracts

P-1. Hydrozyl Impurities in Quartz Crystals and Their Radiation-Induced Dynamics
Harish Bahadur, National Physical Laboratory, India

In the present day technology, crystalline quartz is used in a variety of electronic devices including crystal oscillator and filters for precision frequency control, clocks for microprocessors, temperature and mass sensors and accelerometers etc. Quartz is grown hydrothermally which results in a variety of hydroxyl defects. Trivalent aluminum is the most pervasive impurity in quartz crystals that substitutionally replaces Si 4+. Other substitutional impurities that a quartz crystal can have are Ge, Fe and Ti etc. The impurity-related point defects get modified when ionizing radiations pass through the bulk of a resonator. The radiation-induced modification of such defects alters of the interatomic forces. This changes the elastic constants and finally the resonance frequency of a crystal resonator. This paper presents our investigations on hydroxyl impurities in a variety of quartz crystals and their radiation-induced dynamics.

We present near infrared absorption measurements (in the region of 3100-3700 cm-1) on quartz crystals to characterize the aluminum- and alkali-related hydroxyl defects in a variety of natural and cultured quartz crystals. Quartz samples were irradiated with electron beam of 1.75 MeV and dose of 2 Mrad at 77 K before and after irradiation at 300 K. While the alkalis in quartz move under irradiation field only if the sample temperature is about or above 200 K, the protons move at all temperatures down to 10 K. Therefore, irradiation at 300 K allows movement of both, protons and alkali ions, thus breaking away the aluminum-alkali centers into a mixture of Al-OH-and Al-hole centers. We have measured the natural quartz crystals with nearly similar Al and widely varying H-levels. For a similar radiation dose at 300 K, contrary to expectation, a lesser number of Al-OH- centers are produced in crystal with higher H-level than the sample with low-H quartz. At the present stage of work, we expect this due to jamming in the kinetics of large number of protons in high-H crystals for steric reasons which prevents them to reach Al-sites after irradiation.

Further, the spectral measurements carried out on a variety of natural and cultured quartz crystals show that besides the conventional growth defect bands in natural as well as cultured material, the presence of a small band at 3595 cm-1 occurs due to the presence of Ti in quartz crystal lattice. Irradiation effects have been reported by irradiating the samples at 77 K before and after 300 K-irradiation .While Ge-doped cultured quartz exhibit the production of some new radiation induced defect bands, the results show that among the two substitutional impurities Ge and Ti in quartz other than the aluminum, the presence of Ti is not as deleterious as Ge. Results have been discussed in terms of fundamental considerations governing the radiation induced defect dynamics of point defects in crystalline quartz.

P-2. International Chart Of The Nuclides - 2001

T.V. Golashvili, V.M. Kupriyanov, A.A. Lbov, A.P. Demidov, Head Scientific Data Center, Central Research Institute of Management, Economics and Information (Atominform), Ministry of the Russian Federation for Atomic Energy, Moscow, Russia
V.P. Chechev, Radionuclide Data Center, Khlopin Radium Institute, St. Petersburg, Russia
Zhao Zhixiang, Zhuang Youxiang, Zhou Chunmei, Huang Xiaolong, China Nuclear Data Center, China Institute of Atomic Energy, China National Nuclear Corporation, Beijing, China
M.S. Antony, Centre de Recherches Nucleaires et Universite Louis Pasteour, Strasbourg, France
Akira Hasegawa, Junichi Katakura, Nuclear Data Center, Japan Atomic Energy Research Institute (JAERI), Tokai, Japan

The new Chart of Nuclides has been developed as the updated International Chart of Nuclides-1998[1)]. It contains brief information on characteristics of all isotopes of 118 chemical elements known by 2002. This Chart of Nuclides is a peculiar "wall guide" on nuclides and intended for being used by wide circle of experts of different level (students, graduate students, engineers, scientific researchers), who would like to have primary true information on stable and radioactive nuclides.

Unlike widespread nuclide charts[2,3,4)] that also bring brief information on nuclides, the present Chart of Nuclides contains EVALUATED values of the main characteristics such as mass excess, nuclide percent abundance, cross sections of thermal neutron induced activation for stable and natural long-lived nuclides; mass excess, half-life, decay energy for radioactive nuclides. These values are supplied with the standard deviations. They have been obtained on the basis of the information from database of Head Scientific Data Centre (Atominform, Moscow) and the Radionuclide Data Centre (RDC) at the V.G.Khlopin Radium Institute (St.-Petersburg) including the evaluated data, presented in the *ENSDF-2000* international file[5)], *Table of Isotopes*[6)] and *Table of Radioactive Isotopes*[7)], as well as their own evaluated data obtained by RDC experts.

The uncertainties of the recommended values are parenthetical and provided with the number of units of the last significant digit of the value: for instance, 40.1(22) means 40.1± 2.2.

Nuclide mass excesses, Δ, are expressed in MeV with $\Delta(^{12}C)=0$ and corresponded to data of reference[8)].

Half-life evaluated values (with uncertainties) are presented for radioactive nuclides. Nuclide percentage in natural mixture of isotopes for a given chemical element is mentioned for stable nuclides instead of half-life. Both values, i.e. half-life and abundance of isotopes in natural mixture, are presented for natural long-lived radioactive nuclides.

Basic decay types with percentage of branching, and evaluated values (with uncertainties) of decay energies (Q-values, in keV) obtained on the basis of data[5,8)] are presented.

Basic types of radiation (particles and photons) and mean values of radiation energy per decay (keV/decay) obtained on the basis of data[7)] and RDC evaluations are presented. Mean radiation energy per decay <R> is a quantitative characteristic indicating the contribution of the given radiation type to the energy (Q) released in the decay.

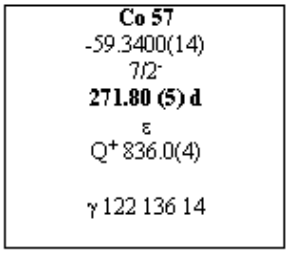
Radiation capture cross sections (in barns) induced by thermal neutrons (activation cross section) are presented for the stable and natural long lived nuclides in accordance with reference[10)]. Also the energies of the most intensive gamma-rays (in keV) are presented.

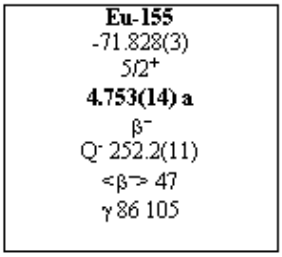
Nuclides in the chart are arranged as Z-N diagram, where Z is the number of protons in a nucleus, N is the number of neutrons. Z grows on along the vertical from bottom to top; N grows on along the horizontal from left to right. The following information for each radioactive nuclide is contained in the information boxes arranged along the lines:

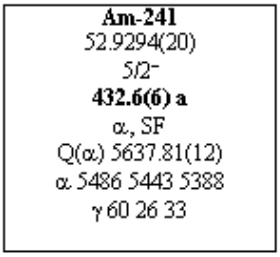
1. Nuclide symbol with mass number;
2. Mass excess;
3. Spin of ground state of nucleus;
4. Half-life;
5. Decay modes;
6. Decay energy;
7. Average radiation energies;
8. Energies of the most intensive radiation components;
9. Thermal neutron activation cross section.

All the values in the first five lines are arranged in such a way that information on the same characteristic for different nuclides is put along the same horizontal line.

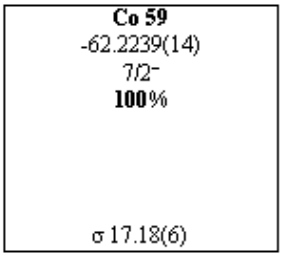
Below the examples of the information box are given for ^{57}Co, ^{155}Eu and ^{241}Am.

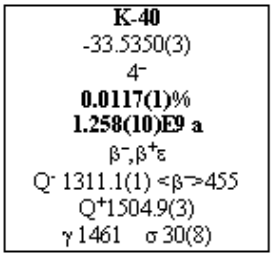






As for the stable nuclides, the abundance of nuclide in natural mixture of isotopes (percentage) is indicated in the forth line and the thermal neutron activation cross section is indicated in the last line. Below the examples of the stable and natural unstable nuclide information boxes are given for ^{59}Co and ^{40}K.





As to history, the necessity to develop the international charts of nuclides was discussed in 1994 at International Conference on Nuclear Data for Science and Technology, Gatlinburg, the USA. IAEA international working group had confirmed that there is a necessity to develop the international charts of nuclides. Opinion of more than 200 respondents from national and international organizations as a result of 1994 - 1996 attitude survey was the reason for developing the international charts of nuclides.

References

1) Zhao Zhixiang, Zhuang Youxiag, Zhou Chunmei, Huang Xiaolong (China), M.S.Antony (France); Akira Hasegawa, Junichi Katakura, (Japan); V.P.Chechev, T.V.Golashvili, A.A.Lbov (Russia). *International Chart of Nuclides-1998*. Scientific Head of the Project: T.V.Golashvili. Atominform, Moscow, 1998.
2) *Chart of the Nuclides*, Knolls Atomic Power Laboratory, Fifteenth Edition, U.S. Department of Energy, 1996.
3) *Chart of the Nuclides*, Nuclear Data Center of JAERI, 2000.
4) M.S Antony, *Chart of the Nuclides* - Strasbourg 1992: Centre de Recherches Nucleaires et Universite Louis Pasteur, Strasbourg, 1993.
5) *Evaluated Nuclear Structure Data File-2000* and *NUDAT*, National Nuclear Data Center, Brookhaven National Laboratory, USA.
6) R.B.Firestone. (Ed.) C.M. Baglin, (CD-Rom Ed.) S.Y. Frank Chu, *Table of Isotopes*, Eighth Edition, 1998 Update, John Wiley and Sons, New York (1998).
7) E Browne, R.B. Firestone, (Ed.) V.S. Shirley, *Table of Radioactive Isotopes*, John Wiley and Sons, New York (1986).
8) G. Audi, A.H. Wapstra, *Nucl.Phys.* **A595** (1995) 409.
9) T.V. Golashvili, V.P. Chechev, A.A. Lbov, Nuclide Guide, Moscow, 1995, Atomniform. P. Raghavan, At. Data Nucl. Data Tables 42 (1989) 189.
10) T.S. Bulanova, A.V. Ignatyuk, A.B. Pashchenko, V.I. Plyaskin. Radiation capture of neutrons. Handbook. M.: Ehnergoatomizdat, Moscow, 1986.

P-3. Electrical Conduction Mechanism in Potassium Boro-Vanadate Iron Glass System

Harshvadan R. Panchal and Dinesh K. Kanchan, The M. S. University of Baroda, India

The Electrical Conduction mechanism of xK2O: (100-x-y)[(1+n)V2O5: B2O3]: yFe2O3 where x=0,5, 10 . . 20, y= 5, 7.5, 10, 12.5 and 15 and n = 0.2 to1 in step of 0.2 glasses was explained on the basis of Mott's theory. The dc conductivity measurements of present glass system were carried out in the temperature range of 315-433 K for all different glass compositions. The decrease of the conductivity and increase in activation energy has been observed with the increase of Fe2O3 concentration. Estimated small polaron radius was to be smaller than the atomic site spacing (V-V spacing) and greater than the radius of iron on which the electron is localized. Present glass system exhibit a semiconducting adiabatic hopping due to a small polaron.

P-4. Database of Korean Mushrooms

Duck-Hyun Cho, Won-Kyung Cho and Jae-Yon Chung
Division of Life and Technology, Woosuk University, Korea

Hyung-Seon Park, Bu-Young Ahn and Kang-Ryul Shon
Bio-Resources Informatics Department, Korea Institute of Science & Technology Information, Korea

Korea has a good environment for development of mushrooms. In summer, there are a lot of rain and high temperatures. Also, it makes possible to grow heavy forests consisting of needle and broad-leaved forests. About 2000 species of mushroom have been studied from basdiomyctes to ascomycetes. Among them, 1,500 species are constructed with database. The mushrooms (higher fungi) are an important part in ecosystem as a decomposer taking responsible for recycling materials. From ancient times, mushrooms have been broadly used in food sources, pharmacy and forests resources. However, many living things have been endangered by environmental pollution and ecological destruction. The higher fungi also are not an exception. This database contains items of mushroom

(higher fungi) from Korea according to the classification: species, genus, family, order, class and division; to the application: pharmaceutical purpose, food sources, culture, toxic, anti-cancer; to the ecological resources: symbiosis, rotten tree; to the geographical distribution and to the illustrated literature. Information retrieval system is also available using KRISTAL? for query searches on the Web in URL http://ruby.kisti.re.kr/~mushroom

---

P-5. Reference Database of Korean Insect Diversity
Soo-won Cho, Chungbuk National University, Korea
Hyung-Seon Park, Korea Institute of Science & Technology Information, Korea

Unlike most of sciences, the field of taxonomy often requires information on species described for the first time. This means the older the paper is, the more important its significance on species description is while more difficult and more expensive it is to get. The Internet offers many useful reference databases in many fields of sciences, but it is neither easy to find a good reference database on systematic nor enough to find specific names listed in the paper by searching abstracts. We are currently building a new database on Korean land arthropods for the following reasons and advantages. First, it is all written in English for enhancing international usage. Second, it can be searched either by reference or by taxonomy, and each paper lists species names listed, with further useful information such as type information on new species, illustrated species images in the paper, and even GenBank data information. Third, we built it for the Internet search by using php and mySQL. Although the project needs a few more years of work in depth and need regular updates, we believe its utility for insect systematists is very high, especially for a preliminary study for a new taxonomic work.

P-6. Cooperative Double Blind Study of Pseudomonads and Related Organisms
Micah Krichevsky (Presenter), Bionomics International, USA
Paul De Vos, University of Gent, Belgium
Surang Dejsirilert, National Institute of Health, Thailand
Deborah Henry, University of British Columbia, Research Centre, Canada
Jorge Lalucat, Universitat de les Illes Balears, Spain
Edward Moore, The Macaulay Research Institute, Scotland, U.K.
Masoumeh Sikaroodi, George Mason University, Prince William Campus, USA
Jane Tang, American Type Culture Collection, USA
Sue Whitehead, Children's & Women's Hospital of BC, Canada
Hans Yu, Health Canada, Biotechnology Section, Canada
Yuguang Zhou, CAS, China General Microbiological Culture Collection Center, Institute of Microbiolgy, China

"The identification of an organism and data to substantiate its identity is one of the critical building blocks that form the basis for risk-based assessment of biotechnology substances Commercial claims to the contrary, microorganism identification is far from trivial. Proper use of identification techniques may require rigorous analysis and in some circumstances may still be problematic due to difficulties and limitations of the science, and the nature of microorganisms themselves." Operational Plan for OECD Guidance Document: The use of taxonomy in the risk assessment of micro-organisms (Segal and Yu, 2001).

Sponsored by Health Canada, we initiated an international double blind study of isolates of two bacterial genera important in biotechnology, Pseudomonas and Burkholderia. One goal is databases to support regulatory deliberations. The databases will associate genera, species, pathovars, serovars, difficult to identify isolates, etc. The primary observations describing the groups and relationships comprise the base data. Analyses yield information on: 1) relationships among known species and hard to identify isolates (of both medical and ecological interest), 2) consistency of characteristics overall and within groups, and 3) a median set of characteristics for further analyses.

The study is not a taxonomic study, per se. Rather, it is disclosure of the consistency (or lack thereof) of the ability to identify these bacteria by laboratories with differing missions and expertise. All participating laboratories have extensive experience with their methodologies and the organisms in question. The participating laboratories vary widely in mission and location.

The central coordinating laboratory (CCC) distributes the cultures to the seven laboratories labelled with unique random numbers. The results are collected and analyzed by the Principal Investigator (PI). The CCC and the PI are independent of the data generation. The data are normalized using a standardized controlled vocabulary and imported into a single object-oriented data management system.

The main aims of the study are to:

1. Ascertain ability to identify to "species" or below
2. Find features robust to error and differing methodologies
3. Delineate useful methods of identification protocols

Not among the aims are:

1. Compare (evaluate) laboratories
2. Standardize test procedures

P-7. Factual Database of Native Flora Seeds in the Korean Peninsula

Hong-Gi Jang, Sang-Uk Chon, Byoung-Sik Pyo & Sun-Min Kim
Biotechnology Industrialization Center, Dongshin University, Korea
Sook-Young Lee, The Institute of Plant Genetic Resources, Dongshin University, Korea
Hyung-Seon Park, Korea Institute of Science and Technology Information, Korea

Each combination of climate, soil and topography has its own characteristic type of plants and each area has its own unique plant species. Korea has a humid, temperate, East Asian monsoonal climate with rainfall heavier in summer than winter, and is geo-morphologically characterized by abundant hills and mountains, which occupy 70 % of its territory. Facing with knowledge-based and technology-oriented society, construction of infrastructure for valuable genetic resources from plant seeds would be the key of biotechnology in the 21st century. Seeds from 600 common plants including weeds, crop plants, and herbs and resources plants were collected in the southwestern part of the Korean Peninsula. Seeds collected were carefully stored in refrigerator until use, investigated morphologically, and photographed. The seeds in database were described with color photographs, their taxonomical position, and morphological characteristics. Korean-English bilingual description of the species included Korean name, family, scientific name, English and Japanese common names, habitat, biotechnological importance, distribution, propagation and characters in eco-physiology and keys of correct identification of each plant part such as leaves, stems, roots, fruits and seeds. In describing plant species, difficulties also arise from the variation that occurs within species, depending on where the plant grows under natural or agricultural conditions. Database was converted from MySQL and constructed using a PHP (http://ruby.kisti.re.kr/~seeds).

Key words: Factual database, native flora seeds, plant genetic resources

P-8. Construction of Antibody Database
Dong-Jun Lee & Chan-Seob Shim, ELPIS-BIOTECH, PaiChai University, Korea
Hyung-Seon Park, Korea Institute of Science & Technology Information (KISTI), Korea

A number of genome maps has already been completed, and a working draft of entire human genome project was announced years ago. The proteomics is on going study and priority goes to the functional genomic and analysis. Antibody is the one of key factors in immune system, and the importance has given to the researcher's activities on functional analysis of protein. More than 700,000 antibodies are available for the research up to now and more antibodies will be produced in proportion to the rapid growth of biology. Antibodies play an important role in a variety of research areas including biotechnology, medicine and diagnostics. Antibody database has built and the homepage is available at http://ruby.kisti.re.kr/~antibody/english/index.html. This site contained about 3,000 human oncogenes out of 6,000. Totally up to 430,000 antibodies that human have will be accumulated. At the same time, the antibody data including mouse, rabbit and sheep will be stored, respectively. Besides, the basis of epitope sequence analysis using by bioinformatics will be constructed. The main purpose for the construction of database is to provide possible information for new medicine drug design and immuno-chip development for new protein screening and so on forth. This antibody database will have a great effect on deciding future course of protein research.

P-9: Database Construction for the Collective List of Descriptions of Bacterial Isolates from Korea
Oh Hyoung Lee, Department of Biology, Mokpo National University, Korea
Hyung Dae Koh, Department of Multimedia Engineering, Mokpo National University, Korea
Soon Young Park, Department of Electronics Engineering, Mokpo National University, Korea
Hyung-Seon Park, Korea Institute of Science & Technology Information, Korea
Kye Jun Lee, Korea Institute of Science & Technology Information, Korea

Bacteria are commonly used as tools or materials for productions of industrial substances such as antibiotics, specific enzymes or hormones, and hence comprising the base of high-valued biotechnology. For these reasons, bacteria have been the main targets for worldwide biodiversity competitions to pre-occupy them, and the efforts to find out new bacterial strains are still made all over the world. But the chance to isolate and identify a new bacterium is getting more difficult to obtain nowadays, because not only the process is complicated but also there is huge information to know about the existing bacteria to compare with the coming new one. For the challenging treasure hunter the need for easy way to grasp and figure out all these information is, therefore, beyond description. This is the reason why we are undertaking to construct database for the collective list of bacteria so far isolated from Korea. Here we constructed a computer-aided program to make a formal, comprehensive description about a given strain. It includes the name of strain, the information about the isolation process such as the time, source, purpose, and methods, the general descriptions about morphological, physiological, cultural, biochemical, and molecular biological characteristics of the target microbe, its specific role or function in nature, and the literature site that describes about these.

So far, 348 bacterial strains have been described in the database, and they are being displayed through the Internet with the URL, http://ruby.kisti.re.kr/~microb

P-10. Factual Database of Noctuid Moths (Insecta, Lepidoptera) in Korea
K.T. Park, J.S. Lee & J.Y.Kim, Center for Insect Systematics, Kangwon Nat. Univ. Chuncheon, Korea
K.J. Lee, KISTI, Daejeon, Korea

The insect fauna of the Korean peninsula has not been extensively explored, due to lacks of specialists, representing less than 15,000 species. The first attempt for the factual database of insects in Korea was initiated by KISTI, for some well known groups of Lepidoptera from 2000, including butterflies, noctuid moths, and leaf-rollers. The structure of the database was established with 14 fields, representing general specific informations, including images of adult, taxonomic status, specific bibliography, biological information including collecting data and hosts, worldwide distributional, and etc.

From this year, the project will be focused on the noctuids of Korea (about 1,000 known species), with more detailed informations which can be used more effectively to researchers and general users. All the previously known localities will be transferred into GIS system, and some new informations on localities in North Korea will be added. Also illustrations of the genital figures of all available species will be given for the key of the specific identifications. These informations will be also very helpful guidance for the taxonomic and zoogeographical study, not only for Korean but also for foreign researchers.

P-11. The Date Conversion DB Between Luni-Solar and Solar Calendar in Korea
Young Sook Ahn, Bo Sik Han, Kyung Jin Sim & Du Jong Song, Korea Astronomy Observatory, Korea

Bu Young Ahn, KISTI: Korea Institute of Science and Technology Information, Korea
We arrange Korean ancient calendar with Solar calendar day from Koryeo dynasty to Choseon Dynasty(A.D. 918-1910). In this period, we have two representable history books and several books, and most of information for date are found from them, Koryeo-sa(1) and Choseon Wangjosillok(2), etc. In those books many astronomical data and calendar information data are contained. Most of the Luni-solar dates can be converted to the Gregorian dates straightforwardly. But uncertain data are identified and converted with solar eclipses, historical events and lunar phase - calculation etc.

We find that arranged chronological tables during Koryeo and Choseon Dynasty are somewhat different from those of China and Japan. In addition, we calculate the Solar calendar and Luni-solar calendar during 1900-2050 using DE200 package.

Now we construct the database system with above data, during A.D. 918 - 2050 and many people will get information of the calendar date which they want using our DB system based on Internet.

(1) Koryeo-sa : The annals of Koryeo dynasty(A.D. 918- 1392) in KOREA.
(2) Choseon Wangjosillok: The annals of the Choseon dynasty(A.D. 1392-1910) in KOREA.

P-12. Algorithm "Skeleton, Segments, Trace" (SST) for digitizing the analog geophysical records
A. Burtsev, Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia
A.D.-Gvishiani, Director of the Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia
M.N.-Zhizhin, Center of Geophysical Data Studies and Telematics Applications IPE RAS, Russia

There are a lot of essential seismic records had been fixed on analog medium but they cannot be used in automatic analysis procedures, because ones require time series. To solve the problem we have developed mathematical algorithm SST and applied it to create graphic application to transfer analog records to the digital time series. The algorithm combines five stages: image quantization, skeletonizing, segmentation, building trace, and the last one, interpolation retrieved trajectory to correspond it with physical measurement. First stage covered all preliminary image processing to make it ready for skeletonizing. To build image skeleton and retrieve linear structure methods of mathematical morphology or distance transform are used. Then nodes of the skeleton are removed to have a set of primitives. To build trace we select and join linear structure covered the trace of the analog recorder pen (dynamic programming is used to find required primitives and set their order).

P-13. The Mackenzie GEWEX Study Data Archive: An Enhanced Dataset for Climate modelling
Robert W. Crawford, Meteorological Service of Canada, Canada

The Mackenzie GEWEX Study (MAGS) is aimed at improving our understanding and prediction of the role of water and energy cycles in the climatic system in the Mackenzie River basin. The goals of the MAGS Data Management system are to establish, maintain, describe, and promote accessibility and distribution of the data sets necessary to meet the MAGS objectives in the data sparse Mackenzie River basin. These goals are being met by MAGS in a number of ways: through the availability of data and information through the World-Wide Web; through the production of specialized data products; and through the development of policies to govern data exchange and participant interactions.

The MAGS web site is the primary method of providing information on the activities and data collected in the project for its many participants in universities and government offices across Canada as well as for the outside world. Visitors to the site have access to over 300 pages of information describing the objectives, background, status, and clients for the project. In addition, nearly 1 Gb of data collected in the study area is available through the site.

Additional data is available to MAGS participants through the "Participant's Only" section. Selected data sets are available there in near real time. These include satellite imagery and enhanced observations from the MAGS surface sites and buoys. The GOES imagery is received every 12 hours and the AVHRR data is received within 30 minutes of capture. The enhanced data sets from the surface sites are transmitted daily and contain enhanced temporal resolution data (for example, 15 minute pressure measurements) and non-standard measurements (such as soil temperatures).

Special CD-ROM archives of the datasets collected during a specific case studies and about scientifically interesting processes within the basin are also being produced by MAGS. These archives will provide a lasting resource for future climate change studies.

This paper will describe the contents, structure and utility of the MAGS Data Archive in conducting multi-disciplinary climate research.

P-14. Background Radiation Level at Tinbhangle (Nepal)
B. R. Shah, Royal Nepal Academy of Science and Technology, Kathmandu, Nepal
H.Rahaman, Department of Mines and Geology, Kathmandu, Nepal

A radiometric survey was carried out to measure the background radiation level in the Eastern and Central part of Nepal in order to prepare radiation map of Nepal.The paper presents survey data taken with two equipments LB 1200 and GAD-1. The dose rates in air at the most places vary in the region of 13-20uR/h while they were found from 50uR/h to 1mR/h at Tinbhangle location. Moreover, the differential counts recorded for Uranium, Thorium, Potassium and Total Count were found 90-145, 0-10,80-130 and 7000-9000 cps. respectively. This preliminary radiometric survey clearly reveals that the annual exposure at the bed rock of Tinbhangle location is around 800 mR, which is fairly a high value in comparison to other places stated elsewhere. In addition to high doserates , a higher count of uranium among other radionuclides provide the greater possibility of uraniferous rock.

P-15. Refusing to Participate in Research: Reasons, Study Characteristics and Patient Demographics
Gina Capretta, McMaster University, Canada
Suzette Salama, Henderson General Hospital, Canada

Objectives: Strategies for recruiting patients in clinical trials are frequently discussed in the literature. This study explores the reasons why patients refuse to participate in research and how those reasons associate with patient demographics (age, gender, job status) and study characteristics (clinical area and study parameters).

Materials and Methods: A questionnaire detailing study parameters, patient demographics, and reasons for refusal was completed by a sample of researchers involved in the recruitment process. Researchers and patients were derived from five clinical areas of Hamilton Health Sciences: Cardiology, Endocrinology, Geriatrics, Infectious Disease, and General Research. Description frequency analysis was performed on the results.

Results: 10 researchers completed the questionnaire for 102 patients. 50.9% of patients were male; 49.1% were female. Mean patient age was 59.7 years (SD = 19.2). Overall, not willing to participate in any research at all(20.6%), inconvenient procedures(17.6%), and fear of side effects(11.8%) were the most common reasons for refusing to participate in research. By clinical area, inconvenient procedures was the most common refusal reason for both Endocrinology (17.6%) and Infectious Disease(30.8%) studies. Fear of side effects was most associated with Geriatrics studies(26.7%). Blood sample, urine sample, and medication were the study parameters most associated with fear of side effects, fear of privacy invasion, and time consumption; questionnaire completion was most associated with language barrier. Inconvenient procedures was among the top two refusal reasons among patients aged 40-59 years and was the most common reason among employed patients. Among the elderly (70+ years) and retired refusers, not willing to participate in any research at all was the most common reason for refusal. Patient gender data was too variable to associate with any specific refusal reason.

Conclusions: Among refusers, there is a general lack of interest to participate in research studies. Our findings also suggest specific reasons for refusal can be associated with particular study characteristics and patient demographics. Though more rigorous analysis of these associations is needed in a larger scale study, addressing these participation barriers is important to improving the outcome of patient recruitment.

Recommendations: Increase patient and community education regarding the role of research and encourage researchers to understand reasons why patients refuse participation.

P-16. Multimedia Data Processing and Construction of Database for Ancient Astronomical Heritages of Korea

Yong-Sam LEE and Min-Soo Lee, Chungbuk National University, Korea
Sang-Hyuk Kim, Institute of Classical Korean Science, Korea
Yong-Bok Lee, Seoul National University of Education, Korea
Bu-Young Ahn, Korea Institute of Science and Technology Information, Korea

We have systematically constructed web site on the heritages and cultural traditions in history of Korean astronomy. The site compose of six chapters on ancient astronomy of Korea which are history, philosophy and thought, instruments, calendar, publications, and astronomers. Most of materials contains a lot of pictures, illustrations, and astronomical records in ancient Korea as a data base. Especially users can understand the function and structure of the astronomical instruments using 3-D animation. It would be learned ancient astronomy for users using multimedia database system. The purpose of site is to introduce the history of astronomy to common user and to offer specific materials for research to expert. And also we present a proto-type of Cyber Virtual Museum which will be complete in 2002.

P-17. Construction of Factual Database Based Virtual Science Museum

Bu-Young Ahn, Hyung-Seon Park, Ji-Young Kim & Kang-Ryul Shon
Korea Institute of Science & Technology Information, Korea

KISTI provides the factual information, mostly indigenous data to Korea, through the web since 1994. There are over 25 databases such as biodiversity related data, inherent- domestic mine data, and specialized data includes chemical product/safety and thermo-physical property of Korean Standard material. Those are all classified into three categories to life science, earth science and industrial. Virtual science museum has constructed based upon those databases using VR Panorama technology, and opened to public including professionals who is provided the genuine data. The method applied was firstly, Real and 3D rendering image based VR Panorama; Secondly, VR Object Format; lastly, the Interactive and Immersive Virtual Reality to experience the virtual space. The virtual science museum consists of 5 pavilions in the following 16 themes; Biodiversity pavilion (freshwater fish, coastal fish, mushrooms, insects, domestic plant, indigenous plant, seeds, birds), Fossil pavilion (Korean fossil, fossil animation, period classification), Shellfish pavilion (Korean, World, Rare), Astronomy pavilion (ancient, virtual solar system) and Agriculture pavilion.

P-18. Enabling Collaborative Science Communities Through Data Interoperability

Daniel Crichton and J. Steven Hughes, Jet Propulsion Laboratory, California Institute of Technology
Gregory Downing and Sudhir Srivastava, National Cancer Institute, National Institutes of Health, USA

Advances in computing technologies is providing new opportunities for science research through data sharing. Increasing volumes of data captured in independent data repositories is proving useful to unlock and share within these communities. The Jet Propulsion Laboratory has been working with both the space science community and the biomedical research community to deploy a data management infrastructure that enables interoperability across geographically disparate data systems located at key research institutions within the United States. In September 2000, JPL and the National Institutes of Health (NIH) signed an interagency agreement to explore infusion of space science data systems architectures into a biomedical research infrastructure. One of the principal findings was the similarity in the approach to building a collaborative data management infrastructure for both communities. A key to building this infrastructure has been the development of an architectural framework named the Object Oriented Data Technology framework (OODT). In particular, the OODT architecture decouples the data architecture from the technology architecture and has been adopted as the underlying architecture to support key scientific networks within both disciplines. OODT's technology architecture is based on XML and messaging services, providing a secure infrastructure for exchanging data across the Internet. Each community has developed a comprehensive data

dictionary that has allowed the technology infrastructure of OODT to be driven by the specific semantic implementation. The data dictionary efforts have provided a common language for sharing scientific data sets, and mapping efforts have been conducted at participating institutions to link geographically distributed databases at these institutions together. In addition, both communities have successfully established cross-disciplinary teams consisting of technical, scientific and administrative personnel that have aided in building a solution that is useful to the community. This paper will discuss the technical approaches to building a data sharing architecture along with the similarities between the two scientific communities including a discussion on the technical, policy, and cultural decisions that were made to enable successful deployment and advocacy by the scientific research communities. Finally, it will discuss how the data sharing architecture described supports the collaborative science goals of two key science networks: the National Aeronautics and Space Administration's (NASA) Planetary Data System and the National Cancer Institute's (NCI) Early Detection Research Network.

P-19. New Welding Information System on Internet
Mitsunae FUJITA, Takayoshi KASUGAI Akira OKADA and Junichi KINUGAWA
National Institute for Materials Science (NIMS), Japan

In recent years, a rapid advance has been made in the branch of information processing technology using networks and computers. It enables anyone to transmit valuable information through the Internet, and thus, to play an active role in each world. In the technical field of welding, any information of its theories and that of our experiences in the past, when they are organized systematically in a certain database system, and also, when such a system is widely opened to public on the Internet for utilization by many people, can undoubtedly promote successive transfers and development of welding technology.

National Institute for Materials Science (NIMS) has been constructing a new system for predicting microstructures and mechanical properties of welded joins. It combines a database system of continuous cooling transformation diagrams for welding (CCT diagrams) and an expert system for computing weld thermal histories. In addition, this system employs a technique which has been invented in developing another distributed database system named "Data-Free-Way" (http://dfw.nims.go.jp/ or http://inaba.nims.go.jp/) for advanced nuclear materials and those obtained through some programs of welding research at NIMS in the past.

This paper describes the present state of our new system computing weld thermal histories for predicting properties of welded joints using CCT diagrams database, which is now available through the Internet. Some problems with the database in such a system are also referred to.

P-20. An Integrated Web Resource for CERN's Ecosystem Data
Hua Ouyang, Institute of Geographical Sciences and Natural Resources Research, CAS, China

In order to meet the challenges of understanding and solving the issues on resources and environments at regional or other larger scales, Chinese Ecosystem Research Network (CERN) was established in 1988. CERN consists of 29 stations on agriculture, forest, grassland, lake and bay ecosystems, 5 sub-centers on water, soil, atmosphere, biological and aquatic ecosystems and one synthesis center at present. Under the strong supports of Chinese Academy of Sciences (CAS), with the efforts of over 700 scientists, technicians and managers, CERN has made significant progresses since then. One of these progress is that an integrated, multi-format, high-quality information system has been implemented on the basis of data standardization, database demands analyzing and database design, also an easily accessed WWW-based system has been developed and put into action in the whole network.

The design process of this application system takes following things into account: (1) It should be developed basing on WWW, so each user on Internet can access CERN's data. (2) The data sources should be rich enough to reflect the whole aspects of CERN's stations (3) Different techniques and solutions will be applied for different data sources and different user communities. The data sources of this information system include research and long-term ecological observation data up to now, as well as multimedia data about stations' nature, landscape, society, and

economies, and stations' maps of soil, land use, vegetation and so on, 3D visualization data of stations' landscape. Classified by format, the data involves attribute data, spatial data, text data, image data, sound data, which is individually managed by various kind of software systems, such as Oracle, Arc/info, Visual Foxpro and file systems.

The application system's functions are: (1) data catalog query by keywords or by meta-data. (2) real data browsing through filter conditions. (3) drawing stations' maps and querying attribute data through spatial location. (4) 3D stations' terrain formation and visualization. (5) multimedia realization, including sound play, image loading and viewing, graphic user interface based querying.

Then, how are so many kinds of data being managed and brought into WWW? This paper will give the details involving how to use Oracle products to manage data and develop Web querying pages, how to manage ARC/INFO maps by Oracle DBMS, how to develop JAVA Applet to realize Web GIS functions and multimedia functions.

Nowadays, more and more data are added to this database annually, this system has been accessed by many ecology-related scientific researchers and provided abundant data for scientific researches.

P-21. The Virtual Organization Environment Engineering
Zbigniew Kierzkowski, Poznan University of Technology, Poland

Basic concepts. Elements of virtual organization creation treated as information society technologies - IST determine: A - informational globalization aspects, i.e. methods and techniques of common creation and data resources using in the form of computer files and their network aggregation, B - turbulent surroundings of globalize processes, i.e. rising number of mutual connections and interactions, C - organizational globalization aspects, i.e. features of properties of appearing virtual organization forming by the entity image of information society organization, D - subjective human participation, in functioning organization determined by factors of processes originated or dependent from human, i.e. dependent from intellectual resources, product of clear symbolic reflection - the most human from human activities forms, the most private from private property forms.

Virtual organization information environment development. Through the virtual organization we understand the set of distributed subjects, represented in the virtual environment and co-operating with oneself through global information environment for achieving common profits. In the information environment of the virtual organization factors of human participation and future of the organization property are represented by:

- co-operation and structural properties and organization functions,
- co-operative activities, functional properties and organization tasks,
- group work co-ordination and real processes of dynamically activities organization and
- information flow management for creation of changing virtual organization configurations.

We use existing information systems i.e. homogeneous, autonomous and isolated environments as well as federations of environments i.e. various environments behavior its autonomy but participating in the realization of determined functions of co-operating subjects. The information environment creates multi-layer architecture, communication of consolidated distributed applications representing real processes (inter-organization, mutual co-dependences etc) of virtual organization participants.

Virtual organization environment modeling. Creating virtual systems depend on globalization of activities organization (X, Y), i.e. on internal and external traditional enterprises organizations changes. Internal changes - X depend on competence growth of virtual organization subjects (organizations of separate enterprises and human activities). External changes - Y depend on integration of separate subjects. One can modeling mutual co-dependence of independent variables X, Y characterizing competences many individual enterprises and their compliance in integration within virtual organizations, conditioning with accessibility to information technologies (Z) - IT. Modeling of information globalization factors and organization globalization factors in spaces of parameter vectors: X, Y, Z allow to diagnose nature of creating virtual organizations as phases of structural transformations of traditional enterprises. Structures of new organizations through factors of the organizational globalization (X, Y) and

the information globalization (Z) take features of flexibility and universality. Flexibility is ability to the continuous re-organization. Universality depends on that, the organization models can be used in different subjective domains so in different real socio-technical and economic systems.

Virtual work organization. In virtual organizations the subject and human work forms are changing. The work in the virtual organization distinguishes big degree of co-management.

It is organized around processes. In the virtual environment the process initiates human work activities. One take into consideration the mutual dependence as well as subjective human participation as inter-organizational connections. In bigger degree we take into consideration futures of human behaviors as well as culture of mutual dependences. It is visible new base of functioning (virtual) organization and human work (of virtual work organization), determined as TEAM - Together Everybody Achievement More. Functioning and the work forms in the virtual organization characterize (Z): first - immersion, i.e. degree of surroundings perception intensity as reality, second - navigation, i.e. degree of use know-how of IT tools.

Pattern solutions creation. Considerations refer to data organizations in chosen domains such as access to knowledge, data for science, technology, economic, organization of integrating institutions and enterprises, promotion of development and culture etc. In the virtual organization environment processes of executing particular functions within the organization are represented. The full solution depends on using models of dynamic documents and mechanisms of data management. Dynamic documents and data management mechanisms are models of virtual organization processes. In the virtual organization engineering we use access systems and organized more and more frequently as data and application centers. Run-time communication aspects of creating resources as well as operating data management as main elements of virtual systems engineering are exhibited. Within centers there are realized management mechanisms. Data centers development and synthesis of run-time communication services of data management decides of structural changes of traditional enterprises into virtual organizations. Trends of using solutions concerning communication and interactivity in the access to resources (interactive communication) are appeared. All of it concerns utility features of data centers and the new data organization in the global information environment. They are contribution for designing useable models transforming structurally traditional enterprises in virtual organization structures.

P-22. Metadata Standard Research and Development for the Scientific Databases System
Li Jianhui, Computer Network Information Center, Chinese Academy of Sciences, China

The project of the Scientific Databases(SDB) has been undertaken by the Chinese Academy of Sciences more than 20 years. By now we have built a large amount of scientific databases, which are distributed, heterogeneous and cover many subjects, including chemistry, physics, geosciences, bioscience and so on. In order to share data among them and manage them effectively, and even to provide information service and knowledge discovery in the near future, we started a big project to research and set up the Scientific databases' metadata standard and built a metadata service system.

The Scientific database's metadata standard includes two pars: a metadata framework and a metadata standard set. The Framework defines the basic semantic and syntax rules which suits for all metadata standards of SDB, which can be seen as the metadata about metadata. The metadata standard set is a multi-level architecture, from SDB standard to subject standard to sub-subject standard, just like an Object-Oriented model.

This paper describes the metadata framework and the structure of the metadata standard set in detail.

P-23. Enabling Collaborative Science Communities Through Data Interoperability
Hua Ouyang, Data Management of Chinese Ecosystem Research Network, CAS, China

Chinese Ecosystem Research Network (CERN) has been established to meet the needs of the researches in ecology, environment and resources. It consists of 29 field stations, 5 sub-centers, and 1 synthesis center. One main objective of CERN is to provide soil data, biological data, hydrological data and climate data, which are recorded and collected by field stations, sub-centers and the synthesis center, to the researchers working in CERN and public. Data management is very important for CERN. It relates to how field stations, sub-centers and the synthesis center divide the work and work together to ensure that the high-quality data could be obtained and supplied to the users.

This paper introduces the data management of CERN. The content is listed as follows.

- Data type and data flow - How to divide the data type and how the data are organized.
- Data submission - When and how field stations and sub-centers submit the data.
- Data quality control - Who is responsible for quality control and how to do it.
- Data sharing - What the data sharing policy of CERN is mainly about, including data classification, user classification, data limitation and data access, etc.

Also, the future of data management of CERN will be discussed in this paper.

P-24. Data Base of Research, Science and Technology (dbriptek) in Indonesia
Rukasih Dardjat, Indonesian Institute of Science (LIPI), Indonesia

In 2001, the program of CODATA - ICSU has organized the national database for research, science and technology (DBRIPTEK) in Indonesia.

The Objective of DBRIPTEK will be the facilities of database to get selected information to support policy makers or decision makers, where for analyzing in and the assessment of research project on science and technology policy.

Database was established by cooperation Ministry of state for research science and technology and Indonesian Institute of Science (LIPI).

DataBase of research, science and technology (DBRIPTEK) is established for an integrated files from:

1. Human resources - researchers file
2. Institutions/University files
3. Project of research on science and technology file
4. Publications (Book, journal, article)
5. Patent

Information storage and retrieval system for each file was created for user friendly and it was designed for input, process and output to make statistic of research on science and technology. It can be operated by using microcomputer, Delphi 5.0 software, Microsoft SQL Server 7.0 and window NT and Windows 98.

It will be useful for exchange information with in CODATA - ICSU member or other user by online in network system through internet.

P-25. Establishment and Role of the Database of Scientists and Engineers in CAS
Shuyu Zhao, Chinese Academy of Sciences, China

The Database of Scientists and Engineers in CAS (DSECAS) was established by the Chinese Academy of Science (CAS). The using of DSECAS had made active roles in science and technology management and policy decision in China. It had also made a good role in international exchange in science and technology.

It was about 16700 scientists were selected to DSECAS. They were professors, doctor teachers, and famous contribution scientist in middle or younger age from all of the institutes of CAS. The information of the scientists were added in DSECAS, including the most important contribution, major, title, age, sex etc.. The database was set up by Oracle.

WWW was connected to the database for the internet using. Managers of the institutes in different levels could momentarily modify the database through internet. From the momentarily modification, the database was guaranteed exactly and reliably. Users could get their search results by internet. They could get different information about the database according to the user's levels division. The search results could be individual information of each scientist, or the statistical results according to the user's requirement conditions.

Since DSECAS was established, it had been used in many field. It was useful in science and technology management. It was used for the scientist selection of Science and Technology Meeting, and for the government and local government decision. It was also used in the international exchange in science and technology. It was used for UNESCO to select 100 Chinese scientists to be its international experts, and for the government to select scientist to be the exchange visit scientists etc.

P-26. The use of the relative and absolute models to the calibration of Landsat TM data. Application to the semi-arid land of Laghouat (Algeria)
A.Bensaid, Z.Smahi, T. Iftene, National Centre of Spatial Techniques, Algeria

The precise interpretation and exploitation of remote sensing data is undertaken using radiometric and atmospheric corrections applied to satellite images in order to compensate the effects caused by the observation angle, irradiance and atmospheric conditions. However , a correction model is used to this aim. The mathematical equations integrated in this model allow the improvement of the spatial data quality.

In this paper, two radiometric correction models were used :
 relative model [4];
 absolute model [3].

Therefore, a software package using programme C, was performed. It is now available and used to undertake. It was tested on TM LANDSAT 5 which represent a semi-arid zone of Laghouat (Algeria).

P-27. Provision of the Operating Access to the Legal Information at the Libraries of the Institutions of Higher Education in Uzbekistan
D.R. Yusupov, Uzbekistan

The project of creation of thirteen legal information centres (LIC) at the regional libraries was fulfilled. The aim of the project was peoples legal education. Legal information centres are equipped with modern computer technics and have access to Internet. The legal adviser of legal information centre is a specialist of higher qualification, he possesses the latest information technologies. There are all necessary literature for work in the library, and computer information searching system on the legislation of the Republic of Uzbekistan. This project showed a great demand in legal information, necessity of increasing legal culture for all inhabitants, and first of all the most socially vulnerable parts of society, private employers and students.

The following ways of legal information centre activity were set forth v work with local authorities work with Higher Educational Institutions colleges, schools (secondary) and other educational subdivisions, joint work with NGO (non-governmental organizations) and other public organizations, active work with mass media. A great number of legal consultations and activities set on solution of these problems are accumulated now days.

In order to deepen this project we set forth creation of automatic place of legal adviser, which consists of the following basic parts:

1. Existing informational v searching legal system on the legislation of the Republic of Uzbekistan ;

2. Electric copies of books, documents, magazines of legal and social information;

3. Structured and well v described notes on Internet resources:
   a) on legal, juridical and social information (portals, sites, pages);
   b) forums v discussions , legal consultations;
   c) electronic news-sendings;
   d) centres of distant education of legal and social problems;

4. The basis of inquires itself, consultations and accepted measures, as this legal information centre, so other regional centres. On the basis of these requires collected from the regional LIC, so called Frequently Asked Questions (FAQ v the list of frequently asked questions and answers) will be created. It will let the citizens, who have already applied the LIC get answers for their questions, more correctly draw up theirs inquiries and have definite notion on this centre.

On gathering a great number of inquires an expert system will be created, which will allow to choose consultations on the former questions automatically. The system will let search the information not only under determined inquires, but under the diffuse initial facts. The base of knowledge is formed as productional system with elements of fuzzy logic and technology neural networks, with the help of which decisions in expert system are synthesized.

The main way of solving the problem is fuzzy model compliancy second genus. Expert system lets solve hard formalizing problems, which leading to form (Characteristic event (situation) v Reason v Operation¦, in this occurrence (Inquiry v Legal advice - Result¦. As the jurisprudence concerns to area hard formulizing problems application of expert system with fuzzy logic is the most effective means of acceptance of decisions.

Creation of such automatic legal advisers in LIC and its subdivisions at the libraries of the Institutions of Higher connected with central LIC, will let widen citizen¦s approach to legal information, will make easier lawyer¦s work.

The most important thing is to create valuable base of knowledge and consultations. Basing on this system its realization on internet technologies is necessary. Creations of a network of sites regional LIC with expansion of the user audience and new functionalities:

- constant renovation of legal basis documents of the Republic of Uzbekistan;
- the time-table of their meetings and report on their activity;
- consulting in automatic regime v expert system;
- Internet v forum;
- creation and renovation of frequently asked questions;
- individual consultations by e-mail with the legal adviser;

All these measures will let us explain the idea and contents of legal rules to the citizens in details, holding of seminars, trainings, a great support in the problem citizen¦s legal education. At first the system should be established in Higher Educational Institution libraries, in local authorities bodies and so on. All public libraries, Higher Educational Institutions Libraries, Tashkent State Institute of Law, Library Association of Uzbekistan and Ministry of Justice participate in this project.

P-28. Data Integration in a Data Acquisition System for Material Property Database
Xinyue Huang, Jun Shen and Yongbin Zheng, Beijing Institute of Aeronautical Materials, China

The experimental data of material property from testing laboratory are becoming more and more important. They are widely used in the material property model, the material product quality control and engineering designation. However, it is difficult sometimes to collect these data from labs because the computer platforms are very different. During the past a few decades, they developed their own databases or e-files for their daily testing data. The data format, field name or unit for the same property item could be different. In this paper, the authors tried to develop a data integrate system based ASP. From this system, the data from various sources, such as the e-file of EXCEL and databases of DBASE III and FOXpro, are collected and put into the data warehouse designed for material property based on SQL Server.

P-29. The Computational Methodology of Multiple Data Analysis for Inorganic Origin Oil and Gas--A Case Study in Qiangtang Basin of Tibet
Mingyuan Huo, Institute of Geography and Natural Resources, Chinese Academy of Sciences, China and Assistant to Mayor, Taizhou City, Jiangsu Province, China
Song Chen, Institute of Geography and Natural Resources, Chinese Academy of Sciences, China

The comprehensive evaluation system of oil and gas resources is based on the different types of origins of oil and gas resources. According to their origins, oil and gas resources can be classified into inorganic and organic origins. Researches in oil and gas geology show that the gross amount of oil and gas resources in a certain area is equal to the amount of inorganic origin oil and gas resources plus the amount of organic origin oil and gas resources. From some references, there are some recognized methods to compute the amount of organic origin oil and gas resources, but there are no preceding methods applied to compute the amount of inorganic origin oil and gas resources. Taking Qiangtang Basin in Tibet as an example, this paper firstly proposes a method to compute the amount of inorganic origin oil and gas resources. Based on the data sources, rising velocity (RV), the area of deep-seated structure (ADS), the action time of the earth's crust after massif formation (ATEC), the equivalent of oil and gas (EOG), and the gross dissipation amount of oil and gas (GDAOG),together with the computational model. The result we computed shows that the amount of inorganic origin oil and gas resources in Qiangtang basin covers 40-60 billion tons.

P-30. Scientific Database and Its Application System of CAS
Xiao Yun, Computer Network Information Center, Chinese Academy of Sciences, China

The Scientific Databases is a large comprehensive scientific information service system built jointly by dozens of subsidiary institutes under the Chinese Academy of Sciences(CAS) in the past ten years. It is the first large-scale progrmatic database information system, multi-discipline, multi-type database and available on the Internet. It boasts a multitude of data of China's special resources, offering retrieval service in both concentrated and distributed modes. These valuable scientific resources have greatly enriched the existing domestic and foreign databases. This paper will introduce briefly the project background, current status and the next 5year(2001-2005) plan.

P-31. Ensuring sustainability access to data — value based approach
Conrad Sebego, CSIR/Satellite Applications Centre

One of the most effective ways of collecting primary data quickly is using space borne sensors that cover large areas whilst also providing repetitive geographical coverage. Space borne systems acquire data of the Earth's land, atmosphere and oceans. Satellite derived earth observation data has the potential to support numerous and diverse scientific disciplines, therefore encouraging the sharing, by government, of data reception costs and a concerted effort in focusing research activities in support of government requirements. This is all possible without the developing countries having to carry the major risks of designing, launching and operating a satellite.

Once large databases have been established, more value is realised by extensively mining the data for new techniques resulting in new applications and knowledge. This implies that a data usage policy must be established to govern access to the data, prevent data violations and protect new IP generated by scientists. This paper explores cost effective ways of obtaining primary research data that can be used to create new value by integrating with data from other sources.

The poster emphasises the importance of the value creation chain: i.e. opening the access to raw data with the aim of encouraging generation of new knowledge aligned with government, industrial and indeed scientific challenges.

P-32. ADRES: An online reporting system for veterinary hospitals
P.K. Sidhu and N.K. Dhand, Punjab Agricultural University, India

An animal husbandry department reporting system (ADRES) has been developed for online submission of monthly progress reports of veterinary hospitals. It is a database prepared under Microsoft Access 2000, which has records of all the veterinary hospitals and dispensaries of animal husbandry department, Punjab, India. Every institution has been given a separate ID. The codes for various infectious diseases have been selected according to the codes given by OIE (Office International des Epizooties). In addition to reports about disease occurrence, information can also be recorded for progress of insemination program, animals slaughtered in abattoirs, animals exported to other states and countries, animal welfare camps held and farmer training camps organized etc. Records can be easily compiled on sub-division, district and state basis and reports can be prepared online for submission to Government of India. It is visualized that the system may make the reports submission digital, efficient and accurate. Although, the database has been primarily developed for Punjab State, other states of India and other countries may also easily use it.

P-33. PAU_Epi~AID: A relational database for epidemiological, clinical and laboratory data management
N.K. Dhand, Punjab Agricultural University, India

A veterinary database (Punjab Agricultural University Epidemiological Animal disease Investigation Database, PAU_ Epi~AID) has been developed to meet the requirements of data management during outbreak investigations, monitoring and surveillance, clinical and laboratory investigations. It is based on Microsoft Access 2000 and includes a databank of digitalized information of all states and union territories of India. Information of districts, sub divisions, veterinary institutions and important villages of Punjab (India) has also been incorporated, every unit being represented by an independent numeric code. More than 60 interrelated tables have been prepared for registering information on animal disease outbreaks, farm data viz. housing, feeding, management, past disease history, vaccination history etc. and animal general information, production, reproduction and disease data. Findings of various laboratories such as bacteriology, virology, pathology, parasitology, molecular biology, toxicology, serology etc. can also be documented. Data can be easily entered in simple forms hyper-linked to one another, which allow queries and reports preparation at click of mouse. Flexibility has been provided for additional requirements due to diverse needs. The database may be of immense use in data storage, retrieval and management in epidemiological institutions and veterinary clinics.

P-34. New CODATA Journal
F. J. Smith, The Queen's University of Belfast, Ireland

Earlier this year (2002) CODATA launched its new journal: "Data Science Journal". Details including the aims of the journal and its scope with the first published papers can be consulted on the internet at

http://www.datasciencejournal.org

The first aim of the new journal is that it will be a quality journal , publishing papers about data and databases, but not publishing the data themselves, covering a range of subjects similar to the papers found already at CODATA conferences and workshops. To ensure that the quality of the papers published in the journal meets the standard normally found in other well known international journals, all papers will be refereed by at least two referees. To make the journal available to as wide a range of scientists and engineers as possible in both the developed and developing world, the journal will be primarily an electronic journal accessed over the internet. However, the journal will have the same structure as a printed journal, and after it has been refereed and accepted, each paper will be attributed a set of page numbers with a volume number and year of publishing as in other learned journals. To be successful the first requirement is that the data community of scientists and engineers send data papers to the journal; so I would invite the authors of papers to the Montreal Conference to submit their papers to the journal. Details of how to submit can be found at the above web site.

P-35. A Model for Live Mission Data Systems Using the OAIS Reference Model
Douglas Hughes, Jason Hyon, Sanda Mandutianu, Kathya Zamora
Data Distribution Laboratory, Caltech, Jet Propulsion Laboratory, USA

Space sciences are confronted with an overwhelming volume of data. The data rates are increasing, the granularity of registered observations is continuously refining, and computer technology allows producing terabytes of images and catalogs. The inexpensive emerging storage technologies, combined with the availability of high-speed communications, will offer the infrastructure for extremely large data repositories to be accessible on-line. Mission data will be quickly accessible almost immediately after they have been collected from space observations. On-line science will demand new tools and technologies for data access, data analysis, and data discovery. These trends will enhance the archival operational concepts mainly related to the long-term information preservation, placing an equally important emphasis on rapid data production, and dissemination to consumers.

However, archive storage technologies have lagged almost 5 years compared to real-time storage systems due to issues with media longevity and cost efficiency. Thus, there tend to be separate systems for an on-line processing operation and an archival operation. We do not see dramatic changes in this trend due to the nature of archive requirements in the near future. If there could be a combined system, which satisfies requirements from an operational system and archival system, it would be an ideal system that is cost effective and efficient. In this paper, we are proposing to build an efficient storage system to offer both state-of-the-art storage technology and longevity. We propose to develop a storage system, which combines reliable hardware technology with rigorous system operations concepts. We believe that the proposed system will satisfy both real-time processing system and archive system needs. We will discuss architecture, policies and processes, future trends, and operational concepts for both the hardware and software environment.

The assumption of OAIS (Open Archive Information System) is that information needs long-term preservation. Long term is to be concerned with impact of changing technologies, including support for new media and data formats, or with a changing user community. The near-term preservation is concerned with more stringent access modes, a known and potentially narrower designated community and possibly with real time or online requirements. We advocate architecture at the confluence of these requirements. This model will apply to any space mission data system. The current mission data systems are characteristically built for the duration of the mission, with the main purposes of satisfying the mission needs. We are studying the effects of a longer-term perspective on using and preserving the data, and we determine what is the desirable architecture for achieving both goals.

P-36. Units Markup Language - An XML Schema for Scientific Units
Robert A. Dragoset, Barry N. Taylor, and Michael J. McLay, National Institute of Standards and Technology, USA

The Extensible Markup Language (XML) is the universal format for the exchange of structured documents and data over the Internet. XML is a set of rules, guidelines, conventions, for designing text formats for structured data (e.g. spreadsheets, configuration parameters, financial transactions, technical drawings, scientific data, etc.), in a way that produces files that are easy to generate and read (by a computer), that are unambiguous, and that avoid common pitfalls, such as lack of extensibility, lack of support for internationalization/localization, and platform-dependency. To date, the development of markup languages to address the needs of specific communities (e.g. mathematics, chemistry, materials science, etc.) has either not addressed the issue of encoding measurement units information with numeric data or has addressed this issue independently for each markup language. This poster will include a description of UnitsML (Units Markup Language), a proposed XML schema for encoding measurement units in XML consistent with the SI (International System of Units), and several instance documents illustrating practical use of UnitsML. Adoption of this schema will allow for the unambiguous exchange of numerical data over the Internet. In addition, we will discuss the development of a repository of detailed measurement units information.

Workshops and Tutorials

Saturday, 28 September from 9 AM to 12 PM and 1 PM to 4PM

CODATA Course on Information Visualization

At Delta Centre-Ville Hotel, 777 University Street, Montréal, Québec, Canada H3C 327

Instructors
Nahum Gershon, MITRE, USA
John Dill, Simon Fraser University, USA
Jean-Jacques Royer, CNRS, Nancy, France
Bill Wright, Oculus Info Inc.

The course will give the participants a working knowledge of effective visualization approaches for presenting information and data. Visual representation of information requires merging of data visualization methods, computer graphics, design, and imagination. In contrast with scientific data, i.e. spatially or geometrically based data, information spaces are more abstract and different from physical data spaces and thus require different visualization approaches. The course will cover types of information and visualization of information retrieved from the World Wide Web (browsing and searching), from large document collections, and from databases. The attendees will learn how to make sense of information with visualization. Practical applications will be illustrated using specific case studies. The course highlights the process of producing effective visualizations, making sense of information, taking users' needs into account, and illustrating good practical visualization procedures in specific case studies.

More specifically, the course will cover the following topics:

- What is visualization including examples driving research and development
- Visualization basics
- Perceptual basis of information visualization
- 3D visualization and visualization of spatial data and information
- Case studies including Web visualization methods
- Conclusions and discussion

Course Price:
Regular: 180 US $
Academic Institutions: 120 US $

Including one-day course, course material, break and lunch

Saturday, 28 September from 2 PM to 6 PM.

CODATA Course on Heterogeneous Information Database & Data Warehousing

At Delta Centre-Ville Hotel, 777 University Street, Montréal, Québec, Canada H3C 327

Instructors
Hélène Bestougeff, Université of Marne-La-Vallée, Paris, France
Jean-Jacques Royer, CNRS, Nancy, France

The course will give the participants a working knowledge on managing, exchanging and integrating heterogeneous information in multidimensional databases, 3D modeling, data warehousing and querying tools. Managing heterogeneous database requires a careful design, architectures and techniques for integrating schemas and data into the information system. It requires new concepts and strategies, which deals with developing architectures and techniques for integrating heterogeneous data, with specific techniques to share standard information in mining systems, 3D modeling and web information management. This half-day course will cover the following topics:

- Heterogeneous database integration: Concepts and strategies, which deals with developing architectures and techniques for integrating schemas and data;
- Data Warehousing: Models and architectures which develops new and different data warehouse architectures such as multidimensional databases with related management and querying tools;
- Sharing Information and Knowledge which concerns knowledge management, use of standards, information mining system and web information management;
- 3D modeling of heterogeneous Natural Objects: a tutorial example of implementing a user interface for manipulating heterogeneous 3D complex objects will be illustrated on geological examples.

The course will be illustrated by ongoing developments in formal and experimental work and extended examples will help the auditors to understand underlying concepts and the difficulties with application. The attendees will be able to master the advanced concepts and difficulties in designing and using heterogeneous database information systems. Practical applications will be illustrated using specific case studies. The selected topics will provide scientists, information specialists, engineers, managers, and librarians with new insights in the field of heterogeneous knowledge management systems, including critical aspects of databases decision making, backed by information sharing processes. In addition, managers will find new incentives and materials to support new information tools in their organization.

Course Price: 120 US $

Including half-day course, course material and break

Book: Heterogeneous Information Exchange. Special Tutorial Price
Heterogeneous Information Exchange and Organizational Hubs
Edited by Hélène Bestougeff, University Marne-La-Vallée, Paris, France,
Jacques Emile Dubois, Université Paris VII, France
Bhavani Thuraisingham, The MITRE Corporation, Bedford MA, USA
ISBN 1-4020-0649-7, Kluwer, Academic Publishers, 265p.

See contents at: http://www.wkap.nl/prod/b/1-4020-0649-7

The book will be offered to the attendees at a special reduced price 120 US$ 90 US$

Sunday, 29 September 2002

Environmental Information in Satellite Imagery and Numerical Classification

A CODATA Workshop on September 29, 2002 sponsored by the Task Group on Data Management and Virtual Laboratories and convened by Alexei D. Gvishiani (Russia) and Herbert W. Kroehl (USA)

CODATA Task Group on Data Management and Virtual Laboratories **Agenda for Workshop** **"Environmental Information in Satellite Imagery and Numerical Classification"** **Montreal, Canada** **29 September 2002**	
10h00	Opening remarks : **John Rumble, Herb Kroehl, Alexei Gvishiani**
10h30	**H. Kroehl (NOAA, USA)** Satellite imagery databases at WDC A for Solar Terrestrial Physics and Environmental Studies
11h00	**M. Zhizhin (CGDS, Russia), E. Kihn (NOAA, USA), A. Gvishiani (UIPE, Russia), H. Kroehl (NOAA, USA)** The satellite archive browse and retrieval (SABR) system
11h30	**E. Kihn (NOAA, USA), M. Zhizhin (CGDS, Russia), H. Kroehl NOAA, USA), A. Gvishiani (UIPE, Russia)** The Environmental Scenario Generator
12h00	**A. Kiremidjian, Pooya Sarabandi (Stanford University, USA)** Use of satellite imagery for earthquake hazard and risk assessment
12h30	**J. Bonnin (Louis Pasteur University, Strasbourg, France)** Satellite imagery and natural hazards in Europe
12h30 - 14h30	Lunch
14h30	**A. Gvishiani (UIPE, Russia)** Fuzzy logic based algorithms and satellite imagery studies
15h00	**M. Zgurovsky (KPI, Kiev, Ukraine)** Analysis of ecological state of the Earth by means of satellite information
15h30	**J-O. Dubois (IPGP, Paris, France)** Dynamical systems approach and satellite imagery
16h00 - 17h00	Round table discussion. Conclusions.

An avalanche of data from space-based remote sensing systems is descending on environmental scientists in many disciplines from space physics and ecology, but the number of scientists to analyze the data is not increasing at a similar rate. At the same time environmental data management systems have undergone a dramatic change in order to "mine" information in the archives using intelligent search systems. Can this avalanche of data be managed in a way to assist scientific investigations? Can intelligent systems be designed to extract nuggets of information from the managed archives? What type of image classification information needs to be available from the data system?

Because remote sensing is an indirect measure of discipline specific parameters, the same data are used in applications covering different scientific disciplines. Thus these data need to be managed in a way that supports as many applications and disciplines as possible. For example, the same imagery are used to identify clouds, to detect wildfires, to classify vegetation and to identify auroral features. And sometimes one person's signal is another person's noise; for example, an atmospheric scientist studying clouds will identify many features as clouds including smoke, but an ecologist studying wildfires only wants those clouds identified that obscure the detection of wildfires. Thus clouds should be available to the intelligent search system. What other information should be included and can this information be derived numerically? Can this information be derived numerically? What other atmospheric and space environment characteristics can be automatically characterized?

The workshop will focus on three topics: numerical classification of environmental information in satellite imagery, management of the archives, and intelligent search systems. Workshop presentations will address numerical methods used to identify clouds, aurora, wildfires, snow cover, sea ice, vegetation, cities, land use, environmental change and the effect of natural hazards. It will investigate ways to manage the archives to facilitate the application of these methods. And it will discuss intelligent mining systems that can extract the information that the environmental science community needs to peruse the avalanche of satellite data.

Mathematical techniques to handle and analyze these data will be another topic of the workshop. Modern methods of cluster analysis, pattern recognition and classification with learning will be a focus of the workshop. Fuzzy set and fuzzy logic based algorithms and results of their applications to satellite imagery will be considered and detaily discussed by the workshop. Virtual laboratories tools to handle environmental and satellite imagery data will be discussed at lengh by the workshop.

All CODATA International Conference participants are welcome to take part in the workshop.