Studies on Monitoring and Tracking Genetic Resources

G.M. Garrity, L.M. Thompson, D.W. Ussery, N. Paskin, D. Baker, P. Desmeth, D.E. Schindel and P.S. Ong

Contact information

*George M. Garrity, Sc.D.
Professor, Microbiology & Molecular Genetics
6162 Biomedical & Physical Sciences Bldg.
Michigan State University
East Lansing, MI 48824-4320 USA
email:garrity@msu.edu

Lorraine M. Thompson, Ph.D.
Pair of Docs Consulting
442 Sydney Drive
Saline, MI 48176
email: lorraine@pairofdocs.net

Dave W. Ussery, Ph.D.
Associate Professor, Leader of Comparative Microbial Genomics, Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Kongens Lyngby, DENMARK
email: dave@cbs.dtu.dk

Norman Paskin, Ph.D.
Tertius, ltd
5, Linkside Avenue, Oxford, OX2 8HY UK
email: n.paskin@tertius.ltd.uk

Dwight Baker, Ph. D.
13 Broadview Street
Acton, MA 01720

Philippe Desmeth, Ph. D.
International Cooperation Officer, BCCM-Belgian Coordinated Collections of Micro-organisms, Belgian Science Policy Office, Rue de la Science 8 Wetenschapsstraat, 1000 Brussels BELGIUM
email: philippe.desmeth@belspo.be

David E. Schindel, Ph.D.
Executive Secretary of the Consortium for the Barcode of Life (CBOL), National Museum of Natural History, Smithsonian Institution, P.O. Box 37012, MRC-105 Washington, DC 20013-7012 USA
email: schindeld@si.edu

Perry S. Ong, Ph.D.
Professor and Director,
Institute of Biology, University of the Philippines, Diliman, Quezon City 1101, PHILIPPINES.
email: ongperry@yahoo.com

* corresponding author

EXECUTIVE SUMMARY 1

ABBREVIATIONS AND ACRONYMS 9

INTRODUCTION AND OVERVIEW 12

GENETIC RESOURCES 13

- GENETIC RESOURCE PROVIDERS AND USERS 14
- USE OF IDENTIFIERS IN TRACKING GENETIC RESOURCES 19
- DISCUSSION 24
- REFERENCES 25

ADVANCES IN GENETIC IDENTIFICATION 29

- CURRENT NGS TECHNOLOGIES 32
- NEAR-FUTURE NGS METHODS 35
- NGS METHODS UNDER DEVELOPMENT 36
- DNA BASED METHODS FOR IDENTIFYING GENETIC RESOURCES 38
- ISSUES 40
- FUTURE POSSIBILITIES 41
- DISCUSSION 42
- REFERENCES 44

PERSISTENT IDENTIFIERS 50

- PERSISTENT IDENTIFIER SCHEMES 53
- UNIFORM RESOURCE NAME (URN) 54
- PERSISTENT UNIFORM RESOURCE LOCATOR (PURL) 55
- ARCHIVAL RESOURCE KEY (ARK) 57
- LIFE SCIENCE IDENTIFIERS (LSID) 60
- HANDLE SYSTEM (HANDLE) 62
- DIGITAL OBJECT IDENTIFIER SYSTEM (DOI) 65
- SOME USERS OF PERSISTENT IDENTIFIERS 69
- DISCUSSION 74
- REFERENCES 85

CONCLUSION AND RECOMMENDATIONS 90

EXECUTIVE SUMMARY

Introduction

Technological innovations, in areas such as DNA sequencing and information technology are characterized by exponential development rates and lead to results that are typically unanticipated when first introduced. Three examples demonstrate this clearly. In 1995 it took Fleischmann *et al*. thirteen months to sequence the complete genome of the bacterium, *Haemophilus. influenzae* at a cost of approximately fifty cents per base pair. Today a bacterial genome can be sequenced in less than a day for pennies per base pair and the possibility of sequencing a complete bacterial genome in a few hours for under $1000 looms in the near future. In 1983 TCP/IP, the underlying protocol of the internet, became operational (Internet, 2009). As of June 30, 2008, 1.463 billion people use the Internet according to Internet World Stats (2009) with the greatest growth in usage between 2000-2008 occurring in Africa (1,031.2 %), Latin America/Caribbean (669.3 %) and Asia (406.1 %). On August 6, 1991, the European Organization for Nuclear Research (CERN) publicly announced the new World Wide Web project. Eighteen years later the Indexed Web contains at least 25.9 billion pages (worldwidewebsize, 2009). Today, digital databases and other resources are accessible to anyone anywhere today with an internet connection and a browser on a computer or handheld device, which may be a cell phone.

It is in this environment of rapid technological innovations and global information access in which the Convention on Biological Diversity (CBD) must work to ensure the *sustainable use* of biodiversity as a means to justify and underwrite its preservation. As part of this effort an international regime (IR) on accessing genetic resources and sharing benefits derived from their utilization (Article 15 of the CBD, Access and Benefit Sharing, ABS) is currently being negotiated by the Conference of Parties (COP) of the CBD. The purpose of this paper is to assist the COP in these negotiations by providing a detailed examination of the following technical issues:

> (a) Recent developments in methods to identify genetic resources directly based on DNA sequences;
>
> (b) Identification of different possible ways of tracking and monitoring genetic resources through the use of persistent global unique identifiers (GUIDs), including the practicality, feasibility, costs and benefits of the different options.

Genetic resources

Genetic resources are used worldwide by many different industries, academic institutions and environmental organizations to achieve various goals, ranging from developing new commercial products and processes to exploring new research avenues for cataloging and preserving biotic specimens arising from biodiversity inventories. In Article 2 of the CBD, genetic resources are defined as “genetic material of actual or potential value” and are further defined as “any material of plant, animal, microbial or other origin containing functional units of heredity.” The value of these resources need not be exclusively genetic material. It may also be derived information, such as functional or regulatory pathways, structural polymers or biological functions of an organism that are encoded for by the genetic material, including metabolic products that have some practical applications (*e.g.*, low molecular weight organic acids; anti-microbial agents, such as antibiotics, and other biopharmaceuticals, flavors and fragrances, enzymes for industrial applications).

Establishing provenance of genetic resources and terms of use

Currently, the use of, and access to, specified genetic resources is governed by contractual agreements between the providers and users of those resources. Contractual negotiations that follow the voluntary Bonn Guidelines result in a set of accompanying documents that explicitly detail the terms of any agreement including prior informed consent (PIC) and material transfer agreements (MTAs) and possibly Mutally Agreed Terms (MATs) and Certificates of Origin (CoO). Such documents by themselves do not provide a means by which a specified genetic resource(s) can be tracked, but do establish an important part of the baseline information that must be collected and made accessible to various parties to the agreement. These agreements also establish the conditions for access to both the resources and information over time and should also specify what types of information are required to follow along with any genetic resource and any real or abstract derived products, either for fixed periods of time or in perpetuity. With this minimal information in hand, it becomes possible to devise reasonable and extensible models to track each genetic resource as it moves from its point of origin through one or more user organizations for a variety of purposes.

It should be understood that a large-scale tracking system that meets the needs of the IR does not yet exist. Smaller-scale implementations do, however; and have features that are desirable in the anticipated tracking system for genetic resources. These are discussed in detail in the section ***Use of identifiers in tracking genetic resources***. We have drawn from prior experience with those smaller scale systems to gain useful insights into the requirements of a robust, reliable and trustworthy tracking system that could accommodate the needs of a diverse end-user community working in pure and applied research, international trade, regulation and enforcement. It is important to stress that development of a complete tracking system for genetic resources must consider non-technical issues as well, including realistic policies that address complex social, business, and scientific requirements. This will ensure widespread acceptance and usage. It is not uncommon for technically sound information systems to fail because user needs were not met or the system rigidly modeled practices that became obsolete because of changes in technologies external to the system, but critical to the organizational goals, that were not anticipated or could not be incorporated into the system. This is particularly true in the life sciences and is discussed in the section ***Advances in genetic identification.***

Redefinition of genetic resources and consequences for tracking

Whereas whole organisms or parts of organisms were once the subject of study and trade, contemporary biology has expanded its focus to incorporate molecular and informatics methods (*in silico*). These newer methods allow us to describe living systems not only on the basis of readily observable traits, but also upon their genetic potential based on a direct analysis of selected portions of the genome or the entire genome. As a result, genetic resources are now being used in various forms ranging from extracted DNA (including from mixed populations in metagenomic studies) to various types of sequence data that are stored in public and private databases. These derived genetic resources are readily copied, mobile and readily accessible to a global audience and can be used for a variety of purposes (*e.g.*, expression in heterologous hosts, engineered chimeric pathways, synthetic life forms) that may have not been intended or anticipated in original agreements.

Therefore, it can be argued that rights and obligations under the IR may extend to the exploitation of genetic resources, regardless of how those resources are constituted. Although a discussion of the merits of such thinking is beyond the scope of our charge, we believe it prudent to consider the consequences. Under such an interpretation, a system for tracking genetic resources would have to provide a means for providers to track the uses of the data and information derived from their genetic resources. The task of tracking successive uses of such information, although complex, is theoretically feasible and would require the crafting of appropriate metadata, careful utilization and implementation of a persistent identifier (PID) system and development of custom tracking applications. However, it should also be understood that such a system would have to accurately reflect our current and future knowledge of biology. The vast majority of gene sequences is ubiquitous in nature and oftentimes occurs in distribution patterns that do not necessarily conform to national boundaries. It should also be understood that current technology allows the rapid synthesis and evolution of genes and pathways *in vitro* and *in silico*. Therefore, apparent misuse of a resource by a user or third party may not be actual misuse. Rather, it may be an instance of coincidental use of a like resource obtained independently. It is with these points in mind, that we offer the Secretariat and the COP our observations and recommendations on the agreed upon topics.

Single gene based identification methods

The rapid development of molecular technologies that enables characterization of organisms at a genetic level has opened new possibilities in species identification. In Woese and Fox (1977) produced the first phylogenetic classification of prokaryotes[1] based on the comparison of the nucleotide sequence of the 16S rRNA gene. This gene is universally distributed, highly conserved, evolves very slowly, and plays a key structural role in the ribosome, which in turn is part of the cellular machinery involved in protein synthesis. All life forms, as we know them, possess ribosomes, so according to the early proposals of Pauling and Zukerkandel, the sequence of this molecule could serve as a molecular chronometer, by which the evolution of different species could be traced.

Woese's work revealed that bacteria and archaea formed two deep and very disctict evolutionary lineages. The third lineage, based on this model of evolution, encompasses the eukaryotes (the plants and animals), which characteristically posses a membrane enclosed nucleus and organalles (including the mitochondria and chloroplasts). Eukaryotes possess ribosomes, which in turn contain an 18S rRNA. The eukaryotic 18S rRNA gene shares many homologous regions with the prokaryotic 16S rRNA gene Thus, it is possible to make meaningful comparisons of all species based on the sequence of this gene. Since the sequence of the 16S rRNA gene is approximately 1540 nucleotides in length, there is sufficent information content to allow for very far reaching comparisons.

Woese's discovery has led to a radically different understanding of the evolutionary history of all life, which is generally well accepted and has led to the abandonment of alternative models of classification (e.g., Whittaker's five kindoms). 16S rRNA Sequence

[1] The term prokaryote is a contentious but commonly used name to group bacteria and archaea together based on their absence of a nucleus; a feature found in all eukaryotes

analysis has become the principal method by which bacteria and archeae are now classified. In the past two decades, thousands of new taxa have been described based on this method, along with numerous taxonomic rearrangements. Concurrent improvements in sequenceing methodologies of have greatly accelerated this process. Today, 16S rRNA sequence data is routinely used to presumptively identify bacteria and archaea to the genus level and to deduce community composition in enivornmental surveys and in metagenomic analyses. These efforts are well supported by publicly available tools and highly curated data sets of aligned 16S rRNA (e.g., the Ribosomal II Database, ARB/Sylva project, GreenGenes)

It is now well understood that a single gene may not be adequate to yield an accurate identification to the species or subspecies level and additional gene sequences along with other data may be required. Confounding issues include non-uniform distribution of sequence dissimilarity among different taxa and instances in which multiple copies of the 16S rRNA gene may be present in the same organism that differ by more that 5% sequence dissimilarity. This can lead to different presumptive identifications for the same individual, depending on which 16S rRNA gene is analyzed. We also understand that numerous instances of misidentification and taxonomic synonomies have accumulated prior to the widespread adoption of these methods and that discrepancies between names and correct classification remain to be resolve. In such instances, molecular evidence needs to be used to support taxonomic revision rather than attempting to force-fit earlier concepts into a classification based on reproducible molecular and genomic evidence.

These observations are relevant to the development of a tracking system for genetic resources because taxonomic names are commonly used in the scientific, technical and medical literature as well as in numerous laws and regulations governing commerce, agriculture, public safety and public health. But taxonomic names are not suitable for use as as they are not unique, not persistent and do not exist in a one-to-one relationship with the abstract or concrete objects they identify.

Analogous developments are currently underway in the fields of botany and zoology. Sequence based methods have been applied on a limited basis to various species of eukaryotes for many years. However, it was not until recently that the community began to accept the possiblity that a single gene could be used for identification of eukaryotes. This approach is now being applied in a highly coordinated fashion to build useful resources to identify plants, animals, fungi, protists and other distinct eukaryotic lineages. Consensus is beginning to emerge on a small number of preferred target genes, of a partial sequence of the mitochondrial cytochrome c oxidase subunit I gene is is preferred. This highly coordinated effort is much more recent than the corresponding activities in microbiology, and championed by the Consortium for the BarCode of Life (CBoL) program.

Whole genome sequencing and its impact on tracking genetic resources

In the section ***Advances in genetic identification*** this paper provides an in-depth review of next generation sequencing (NGS) technologies. Because of the rapid pace at which these technologies are evolving this section should be viewed as a set of “snapshots” of the current state of the art, and a harbinger of the future of DNA based identification methods. We discuss methods that are currently in use; those that have just recently become available on the market, (near-future NGS methods); and those that are still under development. These NGS sequencing technologies enable the rapid evaluation of specific regions of the genome of a biological entity to determine to which genus, species, or strain it belongs. (*e.g.,* the *16S* rRNA gene for taxonomic purposes for bacteria; the use of cytochrome *c* oxidase subunit I (*cox1*) for eukaryotes).

Fuelled by innovations in high-throughput DNA sequencing, high-performance computing and bioinformatics, the rate of genomic discovery has grown exponentially. To date, there are more than 500 complete genome sequences and more than 4000 ongoing genome and metagenome sequencing projects covering species ranging from bacteria to yeast to higher eukaryotes. The results that stream forth from these studies is constantly refining and reshaping our understanding of biological systems. As part of the funding requirement of various governmental and non-governmental agencies, the vast majority of these sequences are made publicly available from the INSDC databases (GenBank, DDBJ and EMBL) after brief embargo periods during which time the funding recipients may publish their results. Typically, after one year, the sequence data is open to anyone wanting to publish their own findings or mine those data for other purposes.

All indications are that future genome-based technologies will be “smaller, cheaper, faster”. This will make genome-enabled detection tools available to a wide audience in both developed and developing nations. Clearly, very low cost sequencing technology along with sophisticated bioinformatics tools will soon be available to presumptively identify a genetic resource, with a high degree of accuracy and reliability, at the point of need.

Tracking genetic resources

The concept of identification is central to the goals of the CBD ABS regime, which rests on the fundamental principle that a user is legally obliged to share in the benefits obtained through the use of a particular genetic resource with the provider. Identification is one of the first steps in tracking an item over time. Under some circumstances, identification to the family, genus or species level may be adequate and identification methods based on a single gene may be appropriate (e.g., biotic inventories, wild-life management, ecological studies). However, there is ample evidence based on over half a century of natural product screening and supporting genomic data that such approaches may be inadequate if the trait of interest occurs in subpopulations within a species or is widely distributed across taxonomic boundaries as a result of horizontal gene exchange. A useful tracking system must accurately reflect current knowledge and readily incorporate new knowledge via continuous feedback over a long time frame as transactions involving genetic resources may be long lived (>20 yrs).

The number of items to be identified and tracked within the anticipated system is a challenge and the extent of the task will depend largely on the legally required “granularity” of the identification. Although there is a tendency to view this as a taxonomic problem and the anticipated tracking system as a taxonomic resource, it is decidedly distinct. What is required is a mechanism to track the fate of multiple genetic resources as each is transferred from one party to another and various abstract and concrete products are generated along the way. In some cases the product may be useful for taxonomic purposes and in other cases taxonomic information may be useful for predictive purposes, but in most cases taxonomic information would be ancillary. Systems of such design are challenging as they are open-ended and must work with data of varying granularity. The point is not to define all the types of data *a priori*, but to define lightweight metadata models that define genetic resources and allow them to be permanently bound other to varying amounts and types of information that accumulate about that genetic resource over time. Inherent in such designs are links established through aggregates of foreign keys that may exist within a single system or on a remote systems accessible via the internet.

Persistent identifiers

In their simplest form, persistent identifiers are nothing more than unique labels that are assigned to objects in a one-to-one relationship. Such identifiers are well understood in computing systems and we present examples of identifiers as used in a large-scale laboratory information management system (LIMS) in the section ***Use of identifiers in tracking genetic resources***. When used in the context of the internet, the concept of persistent identification is frequently coupled with the concept of actionablility, implying that the PID is persistently linked to a specific object and when actuated, will always return the same response to the end-user (typically a hyperlink to a specific web page or other form of digital content). In this context PIDs differ from URLs, which are used to create hyperlinks and provide the internet address of where a given object is stored. As the storage location is not persistent, some "behind-the-scenes" mapping of object identifiers to object locations is required (resolution). This topic is covered in more detail in the section ***Persistent identifiers***.

Persistent identifiers are a powerful enabling technology that provides a way to efficiently cope with chronic problems such as broken links and the general difficulty of reliable and reproducible information retrieval on the Internet. For example, PIDs associated with published articles allows rapid and accurate tracking of written works. PIDs are also in use within the life sciences such as the INSDC identifiers (*e.g.*, sequence accession numbers used at GenBank, EMBL, and DNA Database of Japan). However, these are largely institution specific, *i.e.*, used only within the institutions for which they were created, or are controlled by those organizations, such as the PubMed ID, issued by the National Library of Medicine.

Six PID schemes currently used across different domains and by a number of different organizations are reviewed and include: Uniform Resource Name (URN); Persistent Uniform Resource Locator (PURL); Archival Resource Key (ARK); Life Science Identifiers (LSID); Handle System (Handle); Digital Object Identifier System (DOI). This review also addresses the questions that need to be answered when an organization is assessing the need to incorporate a PID scheme into its data management plan.

Each of these identifiers is used in well-defined settings in which the data and metadata models of the underlying repositories were established *a priori*. The identifiers serve as a means of directly accessing a specific record or other form of digital content or the associated metadata. If the identifier is actionable, then it is possible to retrieve the linked object using the familiar interface of a web-browser. However, with the use of web services that provide structured access to the content of interest automatically (*e.g.* from a database or application on a handheld device using embedded PIDs), similar results can be achieved where an interactive interface is not suitable.

An effective and durable PID scheme requires ongoing maintenance and therefore ongoing resources. While some tasks can be automated, responsibility for this ongoing task must be assigned to an agency, program or office or to a trusted third-party who can guarantee reliability and virtually constant up-time to meet the needs of various end-user communities. In the case of integrating a persistent identifier scheme within the ABS process, the use of a trusted third party with the appropriate expertise and resources is probably the best option, especially if that third party is already engaged in such activity for other purposes.

The selection of an appropriate PID for the CBD ABS and related activities will be critical for its broad utility and community acceptance. However, it does not obviate the importance of carefully defining precisely what the identifiers point to, and what will be returned by queries of various types. It is possible to develop a range of PID services that could, for instance, provide a direct link to digital and paper copies of entire documents, such as PICs, MTAs, CoOs and other relevant agreements or permit tracking of genetic resources or parts of genetic resources in a future proof method, or do so on-the-fly. It could also be possible to track the transfer of materials and the corresponding agreements to third parties in a manner that is consistent with the rights and obligations of all parties to the initial agreement or to subsequent agreements. Similarly, the ability to track these genetic resources into the STM, general interest and patent literature is technically feasible.

Services such as these could be facilitated through the use of a trusted third party acting as a clearinghouse for registering ABS-related events (*e.g.,* PIC, MTAs, CoO and other relevant agreements) according to a set of well-understood business rules. With such a clearinghouse in place, it becomes possible to traverse a series of transactions backward and forward in time, even in instances where some ambiguity may exist. By drawing on highly interconnected information, it is possible to follow events, and to accurately recreate those events, when adequate documentation is available. Such a system would be useful for monitoring the use of genetic resources, especially since there will be instances in which long periods of time may exist between the time PICs, MTAs, and CoO are executed and some commercial or non-commercial product results. With the selection of the appropriate PID system a system of this design could support human and machine queries and facilitates the retrieval of all relevant documents from public and private databases, including the STM literature, patent and regulatory databases. This is discussed in more depth in the section ***Persistent Identifier discussion*** (CBD/ABS services)

Conclusion and Recommendations

Reduction to practice will require a commitment of interested parties from different sectors (*e.g.*, government, industries, botanical gardens, museums, academia, etc) to define standards for the key documents that are instrumental to implementing the ABS. Business rules and policies also need to be established in concrete terms so that useful prototypes can be built and assumptions (technical, legal and social) tested and refined. In the ***Conclusions and Recommendations*** section we offer the Secretariat and COP five broad recommendations along with our reasoning. In summary, these are:

1. Promptly establish the minimum information that must be contained in all relevant documents that are required for compliance with the IR (PIC, MTA, MAT, CoO). Stipulate which documents are mandatory and which are optional.

2. Adopt a well-developed and widely used PID system (*e.g*. DOI) that leverages an existing infrastructure and derives support from multiple sources rather than developing a new system or adopting one that is untested in commercial applications.

3. Carefully consider the current and future needs of genetic resource providers and users as the concept of resource tracking is deliberated. Biological and functional diversity of genetic resources are decidedly distinct. The system must be able to accommodate both with priority given to the latter as functional diversity is what leads to practical utility.

4. Deploy light-weight applications that use browser technology for interactive use and publish well documented application program interfaces to support other web service. Develop strong policies governing access and use of the resource to avoid data abuse

5. Deploy one or several prototype tracking systems to validate underlying concepts and refine critical elements that will be needed in a fully operational system.

ABBREVIATIONS AND ACRONYMS

ABIA	The American BioIndustry Alliance
ABS	Access and Benefit Sharing
ALA	Atlas of Living Australia
ARK	Archival Resource Key
ASEAN	Association of South East Asian Nations
ASM	American Society for Microbiology
BGCI	Botanic Gardens Conservation International
BIO	Biotechnology Industry Organization
BRC	Biological Resource Center
CBD	Convention on Biological Diversity
CDL	California Digital Library
CELB	The Center for Environmental Leadership in Business
CENSUS	Census of Marine Life
CGEN	The Genetic Patrimony Management Council of Brazil
CHM	Clearing House Mechanism
CI	Conservation International
CIESIN	Center for International Earth Science Information Network
CIOPORA	International Community of Breeders of Asexually Reproduced Ornamental and Fruit Plants
CITES	Convention on International Trade in Endangered Species
CNRI	Corporation for National Research Initiatives®
COP	Conference of the Parties to the CBD
CSIR	Council for Scientific and Industrial Research in South Africa
CSIRO	Australia's Commonwealth Scientific and Industrial Research Organization
CSOLP	Certificate of source; origin or legal provenance
DNS	Domain Name System
DOI	Digital Object Identifier
EBI	The Energy and Biodiversity Initiative
ENCODE	Encyclopedia of DNA Elements
EoL	Encyclopedia of Life
FAO	Food and Agriculture Organization
FFI	Fauna and Flora International
GAA	Global Amphibian Assessment
GBIF	Global Biodiversity Information Facility
GEF	Global Environment Facility
GHR	Global Handle Registry
GOLD	Genomes OnLine Database
GR	Genetic resources
GRI	Global Reporting Initiative
GRID	Australian Government Genetic Resources Information Database
GSPC	Global Strategy for Plant Conservation
GTI	Global Taxonomy Initiative
GUID	global unique identifier
HMP	Human Microbiome Project
HTTP	Hypertext transfer protocol
ICBG	International Cooperative Biodiversity Groups
ICC	International Chamber of Commerce-the world business organization
IDDRI	Institut du développement durable et des relations internationales
IDF	International DOI Foundation
IFC	International Finance Corporation
IFOAM	International Federation of Organic Agriculture Movements

IFPMA	International Federation of Pharmaceutical Manufacturers and Associations
IISE	International Institute for Species Exploration
INSDC	International Nucleotide Sequence Databases Collaboration
IOPI	International Organization for Plant Information
IP	Intellectual Property
IPR	Intellectual Property Rights
IR	International Regime
ISBN	International Standard Book Number
ISF	International Seed Federation
ISO	International Organization for Standardization
IT	International Treaty (on Plant Genetic Resources for Food and Agriculture)
ITPGRFA	International Treaty on Plant Genetic Resources for Food and Agriculture
IUBS	International Union of Biological Sciences
IUCN	International Union for Conservation of Nature
JBA	Japan Bioindustry Association
LIMS	Laboratory information management systems
LMMC	Likeminded Megadiverse Countries a group of 12 countries located largely in the tropics that have the richest variety of animal and plant species habitats and ecosystems
MabCent	Centre on marine bioactives and drug discovery
MAC	Marine Aquarium Council
MAT	Mutually Agreed Terms
MEA	Multilateral environmental agreements
MOU	Memorandum of Understanding
MSA	Material Supply Agreement
MSC	Marine Stewardship Council
MTA	Material Transfer Agreement
MTDS	Monitoring tracking and documentation system
NAAN	Name Assigning Authority Number
NCRIS	National Collaborative Research Infrastructure Strategy
NGO	Non Governmental Organization
NHGRI	National Human Genome Research Institute
NID	Namespace Identifier
NIH	National Institute of Health
NMAH	Name Mapping Authority Hostport
NPB	Natural Products Branch of the National Cancer Institute
NSS	Namespace Specific String
nts	nucleotides
OAU	Organization of African Unity
OCLC	Online Computer Library Center
PES	Payments for ecosystem services
PGRFA	Plant Genetic Resources used for Food and Agriculture
PIC	Prior Informed Consent
PID	Persistent Identifier
PIIPA	Public Interest Intellectual Property Advisors Inc
PPI	People and Plants International
PSI-Nature	SGKB Structural Genomics Knowledgebase
PURL	Persistent Uniform Resource Locator
RDP	Ribosomal Database Project
RSPO	Roundtable on Sustainable Palm Oil
SABONET	Southern African Botanical Diversity Network
SAI	Sustainable Agriculture Initiative
SAN	The Rainforest Alliance and the Sustainable Agriculture Network
SANBI	South African National Biodiversity Institute

SBA	Sustainable Business Associates
SBS	Sequencing by synthesis
SBSTTA	Subsidiary Body on Scientific Technical and Technological Advice
SCNAT	Swiss Academy of Sciences
SINTEF	The Foundation for Scientific and Industrial Research in Norway
SME	Small and Medium-Sized Enterprise
SMS	Single molecule sequencing
SMTA	Standard Material Transfer Agreement
SNP	single-nucleotide polymorphism
SP2000	Species 2000
STM	International Association of Scientific, Technical & Medical Publishers
TEEB	The Economics of Ecosystems & Biodiversity
TIB	German National Library of Science and Technology
TK	Traditional Knowledge
TKDL	Traditional Knowledge Digital Libraries
TMOIFGR	Tracking and monitoring the international flow of genetic resources
TRIPS	Trade-Related Aspects of Intellectual Property Rights
UNDP	United Nations Development Programme
UNEP	United Nations Environment Programme
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
USCIB	United States Council for International Business
WCS	Wildlife Conservation Society
WGS	Whole Genome Shotgun
WIPO	World Intellectual Property Organization
WSSD	World Summit on Sustainable Development
WTO	World Trade Organization
WWF	World Wildlife Fund for Nature
WWW	World Wide Web

INTRODUCTION AND OVERVIEW

An international regime on access to genetic resources and sharing of the benefit of their utilization (Access and Benefit Sharing, ABS) is currently being negotiated under the framework of the Convention on Biological Diversity (CBD). The fair and equitable sharing of benefits deriving from the utilization of genetic resources is one of the three basic objectives of the Convention and the principles underlying this objective are set out in Article 15 of the Convention. The international regime (IR) currently under development is intended to provide the international framework for the implementation of these key provisions of the CBD, as well as related provisions of the Convention.

At its ninth meeting, held in May 2008, in Bonn, Germany, the Conference of the Parties (COP) to the Convention adopted a calendar for the completion of the negotiations by October 2010 (decision IX/12). The decision provides for three negotiating sessions to be held respectively in April 2009, November 2009 and March 2010. The decision also provides for the organization of three expert meetings to inform the negotiating process. The mandate of the expert meetings is set out in paragraph 11 and annex II of decision IX/12. The expert meetings will focus on: (i) compliance; (ii) concepts, terms, working definitions and sectoral approaches; and (iii) traditional knowledge associated with genetic resources.

With a view to supporting the negotiating process, the Conference of the Parties requested in paragraph 13 of decision IX/12 that the Secretariat commission studies on technical and legal issues, which are central elements of the negotiations, including the following:

(a) Recent developments in methods to identify genetic resources directly based on DNA sequences;

(b) To identify the different possible ways of tracking and monitoring genetic resources through the use of persistent global unique identifiers, including the practicality, feasibility, costs and benefits of the different options;

This study is intended to address those issues.

GENETIC RESOURCES

Genetic resources are used worldwide by different industries, academic institutions and environmental organizations to achieve various goals, ranging from developing new commercial products to exploring new research avenues for collecting and preserving biotic specimens arising from biodiversity inventories, *in situ* or *ex situ* Rapid technological developments in computer sciences, bioinformatics, and biotechnology have greatly expanded the different ways in which these resources can be utilized (Laird and Wynberg, 2008a. 2008b; Parry, 2004).

In Article 2 of the CBD, genetic resources are defined as "genetic material of actual or potential value" and genetic material further defined as "any material of plant, animal, microbial or other origin containing functional units of heredity." The value of genetic resources need not be commercial (*i.e.* have monetary value or be offered for sale or barter), but may be of scientific or academic nature. As the CBD definition also includes the potential value of such resources, the issue that then needs to be addressed is whether most all or all genetic material falls under the provisions of the ABS system. If the former, then the question arises concerning the conditions or characteristics of those genetic resources that are exempt from ABS. Furthermore, the value of these resources need not be exclusively genetic, for example, it may also be derived information, such as functional or regulatory pathways, structural polymers or biological functions of an organism that are encoded for by the genetic material, such as metabolic products that has some practical applications (*e.g.*, low molecular weight organic acids; anti-microbial agents, such as antibiotics, and other pharmaceuticals, flavors and fragrances, enzymes for industrial applications)

Genetic resources are essentially "packets of informational goods" that are presented as biological material (*e.g.*, an entire specimen, a leaf, skin, *etc.*) and include DNA and RNA molecules as well as gene or protein sequences. Some may also regard the end products of gene expression, including proteins and other biopolymers and molecules as genetic resources (Parry, 2004). Each of these elements may have a specific function and potential use and, in some cases, may be subject to specific legal rules, including intellectual property rights. Modern technologies nowadays enable use of the packet as a whole or in isolation and may include use of its component elements. (Muller and Lapena, 2007; Parry, 2004). Therefore, if the ABS benefit-sharing commitment is to be met, providers should share in the benefits (financial or non-financial) that accrue from the commercial or non-commercial exploitation of their genetic resources regardless of how those resources are constituted. Hence, providers must be able to track all the uses that are also made of the bioinformation extracted from the genetic resources they provided to commercial and non-commercial entities. If this reasoning is to be adopted, then it follows that each of these elements must be anticipated and accounted for in any agreement between a provider and a user of a genetic resource. Some reasonable decision must also be made as to which of these components is of sufficient value to warrant such efforts.

Documenting the origin of materials in research and manufacturing is a well understood problem. There are numerous approaches that can serve as examples for considering how genetic resources may be tracked. Typically, these systems are closed, so that only those within a particular organization have access to the information and usually on a need-to-know basis. In most cases, providers of genetic resources have restricted or no access to such systems and therefore would have no knowledge of the use that was made of any material they provided. Likewise users would have no knowledge of other agreements that their providers may have made with third parties. Addressing issues such as rights and obligations of either genetic resource providers or users would be addressed through contractual agreement with compliance largely being an issue of trust.

Genetic resource providers and users

Genetic resource users can be divided into commercial or non-commercial users. The corporate sector plays a dominant role in the commercial use of genetic resources and their derivatives. For example, food and beverage manufacturers use botanical and microbial genetic material to develop compounds to sweeten or fortify food products; many sweeteners are produced using immobilized enzyme technology (bacterial enzymes). Commercial seed companies collect and develop seed varieties for horticulture. Personal care and cosmetics companies research and develop substances to moisturize, color or add fragrance to their products. Future breakthroughs in the pharmaceutical industry may yet depend on the availability of a sufficiently large genetic base (Laird *et al.*, 2008) as a source of new leads, despite the fact that most companies abandoned this route during the last five years.

Non-commercial users, such as botanical gardens, zoos and aquariums, culture collections and Biological Resource Centers (BRC), and academic institutions obtain genetic resources for purposes of preservation, conservation or scientific research. There are times when the line between commercial and non-commercial users is blurred. Examples include, but are not limited to, a botanical garden hosting plant sales or selling seeds; research groups within a university developing processes and/or products (especially at a cellular or molecular level) that may lead to discoveries with potential industrial applications outside the laboratory; and the discovery of small molecules with biopharmaceutical application or biopolymers with practical utility and value as a result of mining sequence, literature, and patent databases for potentially useful enzymes or pathways. Care must be taken that all ABS agreements are respected in carrying out such activities. To complicate matters, there is growing involvement of the private sector in partnerships with public institutions. The increased use of intellectual property instruments to protect innovation in terms of products or processes, also affect who, how and even whether new research can be undertaken. Most importantly, it affects the direction research takes and who controls the results.

Botanical gardens and herbaria play an important role in the conservation of the plant species of the world. As practitioners of *ex situ* conservation these institutions are responsible for the development and maintenance of germplasm collections including seed banks, collections of tissue explants, species recovery programs, and databases. Much of the work of botanical gardens and herbaria depends on access to and exchange of new plant material, which is reflected in the mission statement of Botanic Gardens Conservational International (BGCI): “to mobilize botanic gardens and engage partners in securing plant diversity for the well-being of people and the planet”. The extensive collaboration on developing a global policy on ABS for botanical gardens and herbaria has resulted in the development of a set of non-legally binding Principles on access to genetic resources and benefit-sharing (William *et al.*, 2006).

As with all other aspects of benefit-sharing, institutions need to ensure their internal tracking systems allow them to comply with existing contractual obligations. This is especially true since botanical collections are also increasingly involved in digitisation projects to disseminate data and images of their collections on the internet. This raises the issue of another form of IP that is rarely addressed in the CBD literature: copyrighted material. Most gardens and museums have either captive publishing units or agreements with commercial publishers. This leads to questions concerning the ownership and control of such materials. If released as open access, who decides? Who controls the derivative rights when images and content are republished in field guides and online resources such as the Encyclopedia of Life (EoL)?

Bioprospecting is defined as the exploration of biodiversity for commercially valuable genetic and biochemical resources but should also result in the protection of wild lands and wildlife, through funding of conservation activities. In 2005, the market size for products and applications derived from bioprospecting was estimated to be US$500–800 billion per annum (Christoffersen, 2005). Interestingly, pharmaceutical products represent only about half of the revenues generated from bioprospecting (Parry, 2004). Biotechnology, agriculture, natural products, and food and beverage industries market products derived from genetic and biological resources (Laird and Wynberg, 2008c) and some of these products are less susceptible to market contraction arising through expiration of patent protections.

The emergence of newer, highly efficient techniques in industrial drug discovery, including ultra high throughput screening and combinatorial chemistry, along with too few recent successes by major pharmaceutical companies in natural product screening, are major reasons for the virtual abandonment of natural products recently (Newman *et al.*, 2007; Laird and Wynberg, 2008a, 2008b; Parry, 2004). While of considerable interest to the Parties, it should be understood that in such settings, bioprospecting represented only a small component of much larger interdisciplinary efforts. Samples derived from bioprospecting (along with any potential biotic knowledge) are essentially commodity products that feed into various points in a stochastic process and tracking the outcomes for each and every sample derived from any particular genetic resource is relatively straightforward and routine (Kuo and Garrity, 2002; Newman and Cragg, 2007). Managing the expectations of providers is, however, more difficult as the rate at which meaningful discoveries are made is vanishingly small ($< 10^{-5}$). Managing these interdisciplinary interactions inevitably implies longer timelines compared to research activities driven exclusively by synthetically derived compounds. **Figure 1** is a depiction of the research paths that a company may follow in its efforts to discover commercial benefits from a genetic resource. Finally, legal uncertainties surrounding international agreements governing bioprospecting and biodiversity have led some pharmaceutical companies to terminate their natural product exploration programs (Petersen *et al.,* 2008*)*.

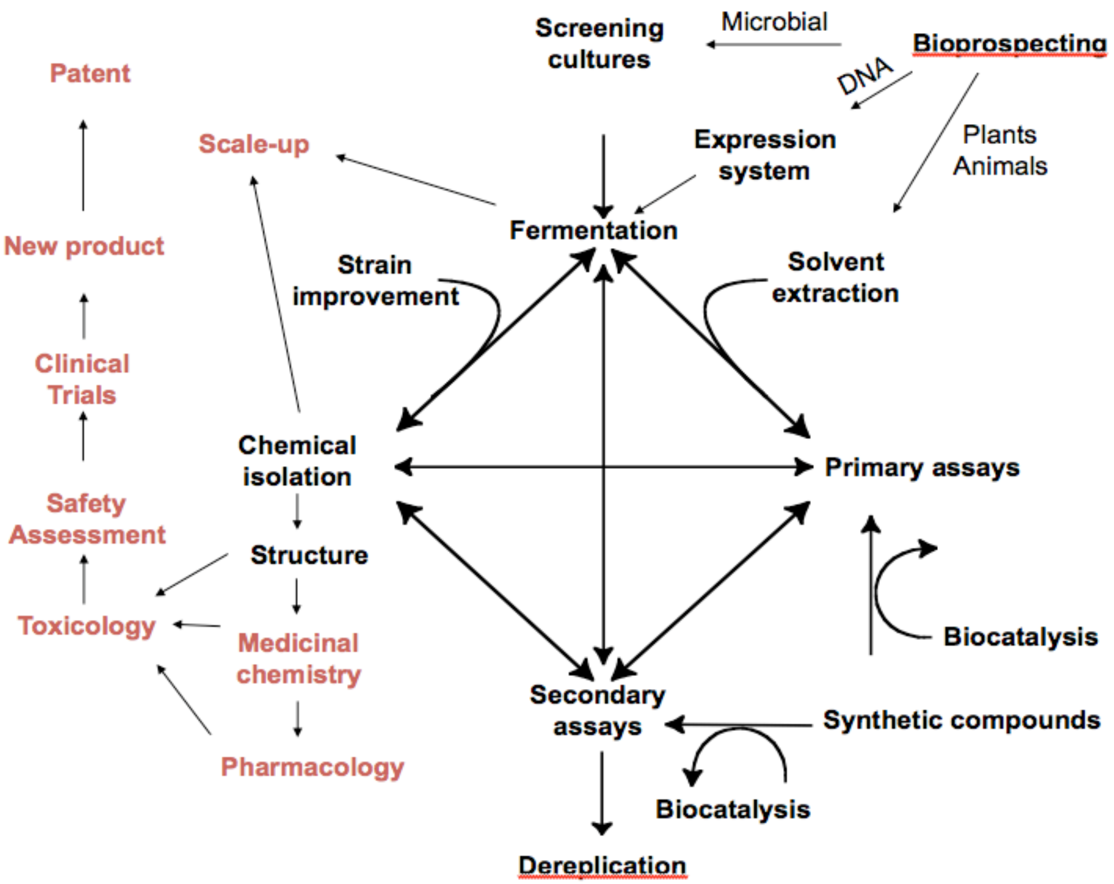

Figure 1 A condensed overview of a typical program used to screen for bioactive compounds such as pharmaceuticals (Based on Kuo and Garrity, 2002). In the "traditional" model of natural products screening, bioprospecting products, consisting of either whole organisms or parts of organisms (plants and animals) are typically prepared as solvent extracts, ranging in polarity and protonation (*e.g.* aqueous, alcohols, ketones, halogenated solvents), dried, and re-suspended in an aqueous form acceptable for screening in batteries of various biochemical and antimicrobial screening assays. In the case of microbial products, the product of bioprospecting is typically a small amount (5 - 10 g) of soil, sediment, leaf litter, tree bark and other plant tissue, herbivore dung, or other similar materials that are likely to harbor bacteria, archaea or fungi. Typically, microorganisms are isolated and purified to prior to screening for bioactive compounds that are produced under batteries of controlled cultivation (often referred to as fermentation, but growth is usually under aerobic rather than anaerobic conditions). Fermentation broths are then screened either directly or after extraction with various solvents used to selectively recover different metabolites that can be concentrated prior to testing. The third product that can be screened in contemporary programs is DNA. Two general approaches are employed. In the first, specific pathways of interest are targeted using probes or PCR primers that permit direct recovery of genes of interest that are then cloned into well-understood heterologous hosts where the expression of those genes can be controlled. As these expression systems are typically bacterial or yeast based, the subsequent steps in the screening process are similar to those for wild-type microorganism recovered from nature. The second approach in use involves extraction of total DNA from either whole organisms or directly from environmental samples (metagenomes) and randomly clone fragments into heterologous hosts for screening.

In the pharmaceutical screening model depicted above, primary and secondary screening assays are centered on "targets" that are relevant in various diseases and act as a mechanism for detecting leads worth follow-up. Primary assays typically filter out >99% of the samples that enter the screening program. Secondary assays are then employed that provide independent confirmatory evidence that a potential lead is worth additional follow-up. Typically, at this step, additional physicochemical data (dereplication) is generated to ensure that the potential lead is not attributed to a known compound. Once there is sufficient evidence that a potential lead is real and reproducible, chemical isolation work begins in earnest. In general, less that 1/1000 screening samples reaches this step. Chemical isolation work typically takes months to have sufficient material in hand to determine the absolute structure of a potential lead compound, and often requires multiple cycles through this "loop", along with many cross collaborations. Once it has been determined that a lead is associated with a novel compound, then small scale studies outside the core program are initiated to begin gathering toxicity and efficacy data. In large scale programs, relatively few leads progress to this stage each year, and even fewer progress to subsequent steps. One could view such programs as a series of exclusionary decision points that a lead must pass through, each addressing different conditions that an acceptable drug must meet, prior to testing in humans.

Historically, microbes have also been screened for biocatalytic properties. Many pharmaceuticals (synthetic and derivatives of naturally occurring compounds) have various functional groups that are amenable to modification to achieve improved properties. Often times such modifications are quite challenging for chemists to accomplish, yet bacteria and fungi possess the capability of making such modifications with high specificity and high yields (typically > 99.999% purity). Thus, wild-type strains recovered in bioprospecting programs can also be used for this purpose.

While the focus of the program above is on biopharmaceuticals, similar approaches are taken when screening for enzymes and other useful metabolites that can be used as chemical feedstock to replace petrochemicals and biofuels. Such programs, however, have a much faster cycle time and do not require most of the downstream research that is required to prove safety and efficacy.

Genetic resources continue to be more attractive candidates for discovery because of the "…limited success of combinatorial chemistry and synthetic compounds over the last decade, limitations to protein engineering, and a realization that natural solutions to the pressures of evolution have resulted in products that could not be engineered in the laboratory" (Laird and Wynberg, 2008a). The ability to isolate DNA directly from samples, without resorting to culturing, also means that the vast genetic diversity of nature can be more readily exploited. Parallel development in bioinformatics and sophisticated molecular tools also mean more detailed information can be extracted from each sample (Newman and Cragg, 2007; Parry, 2004). Baltz (2007) notes however that no new antibiotics have been identified since there has been a shift from "…the traditional methods of identifying antibiotics by screening extracts from actinomycetes and fungi against pathogens" to using "…powerful experimental tools such as genomics, combinatorial chemistry and high-throughput *in vitro* screening". The real limitation in this area comes from a dearth of meaningful screening targets. In the case of human health (*e.g.* chronic disease) this may be attributable to the actual number of privileged structures that are suitable and amenable to screening (about 600). In the case of antibiotics, it is the number of lethal targets in bacteria, fungi, yeasts and viruses that are not present in hosts (Tulp and Bohlin, 2002).

The full impact of these developments on demand for access to genetic resources from high biodiversity areas is still unfolding, but it is likely that nature will continue to be a source for structural leads (also referred to as scaffolds) that will then be modified in the laboratory by a variety of biocatalytic and synthetic approaches. As a result of these advances, microorganisms remain of interest to the pharmaceutical and biotech industries, albeit in different ways than during the peak of natural product screening prior to 2000 (Kuo and Garrity, 2002; Dalton, 2004). New technologies and new approaches have enabled researchers to study previously inaccessible microbes. Biotech companies are searching for microoganisms living in extreme environments such as salt lakes, deserts, caves, hydrothermal vents, and deep seabeds (Wilson and Brimble*,* 2008; Laird and Wynberg, 2008c). The amazing numbers and diversity of microorganisms, combined with their ubiquitous existence, have led to exploring their potential use in everything ranging from energy and water-saving industrial processes, to pollution control and biomaterials (Laird and Wynberg, 2008c).

Metagenomic approaches to bioprospecting provide technical advantages to traditional methods because metagenomic methods can sidestep the often difficult and time-consuming requirements to isolate and cultivate organisms from nature to produce desirable products (Daniel 2004; Riesenfeld *et al.,* 2004). Metagenomic surveys have been utilized to identify novel bioactive chemicals and enzymes of use to the chemical, pharmaceutical and bioprocessing industries (Cowan *et al.,* 2005; Wilkenson and Micklefield 2007; Pham *et al.,* 2007). Early applications of this technology have been found in the identification of novel biocatalysts (Gabor *et al.,* 2004, Lämmle *et al.,* 2007; Li *et al.,* 2008; Riaz *et al.,* 2008; Steele *et al.,* 2009). In most cases, the phylogenetic origin of the genes or biosynthetic pathways encoding desired products is unknown or irrelevant to the applied research. Scientists employing this methodology may find that capturing or maintaining databases of source identifiers is cumbersome and prohibitively costly. Methods have recently been developed to predict source identifiers of metagenomic genes using statistical analysis (Teeling *et al.,* 2004; McHardy *et al.,* 2007). Such analyses may also be capable of predicting both phylogenetic and geographic source because diversity appears to be unique around the world (Fierer *et al*., 2007). If this proves to be universally correct, scientists may opt to predict the sources of their selected useful genes rather than maintain source identifier databases proactively.

Metaproteomics is an emerging field of exploration that goes a step beyond metagenomics and looks at functional diversity of environments rather than genetic diversity (Maron *et al*., 2007). Like metagenomics, metaproteomics offers the opportunity to identify novel enzymes or activities without obligating researchers to identify or culture the biological host (Wilmes and Bond 2006, Noirel *et al.,* 2008).

Synthetic biology is a new field of inquiry made possible by combining biological components (genes, pathways, etc.) from nature, or by creating newly designed biological components fashioned from models identified in nature (Meyer *et al.,* 2007, Keasling 2008, Leonard *et al,.* 2008). New technologies allow scientists to engineer totally novel compounds, enzymes or organisms that seem to have little resemblance to specific organisms in nature. With continuing advances in identifying useful genes or pathways from consensus sequences derived from many organisms, and the ability to quickly synthesize and express nucleotide sequences in model systems, bioprospecting in a traditional sense may be replaced by designed evolution (Channon *et al,.* 2008; Peccoud *et al.* 2008). As the adoption of synthetic biology approaches becomes more widespread, the issues of biodiversity exploitation and intellectual property will become increasingly more complex (Rai and Boyle, 2007; O'Malley *et al.,* 2008).

At the genetic resource level, two key issues for business are related to ABS compliance and intellectual property rights. These topics are intertwined and potentially contentious because of unrealized expectations on the part of both the providers and users of genetic resources. This may be attributed to a lack of transparency, which in turn leads to a lack of trust. Here, a well-designed and properly implemented tracking system could be useful, provided that the complex interrelationships among providers and end-users could be specified, correctly modeled and implemented. It is also essential that such a system not impose undue burden on its users, as that would provide a disincentive. Rather, such a tracking system needs to facilitate interactions between genetic resource users and providers by providing access to new knowledge, information, and data that arises from continual feedback.

In the development of any tracking system of genetic resources it will be necessary to consider tracking the successive uses that are made of such information. Although this makes the tracking more complex, it is in theory entirely feasible through the crafting of appropriate metadata and the careful utilization of unique, persistent identifiers (PIDs) as well as the necessary funding. The technical requirements involved with implementing a well-crafted PID as well as a review of the persistent identifier schemes that are currently most prevalent will be addressed in the section entitled Persistent Identifiers.

Use of identifiers in tracking genetic resources

To understand how PIDs may be used in tracking genetic resources, we believe it may be instructive to revisit the natural product screening process depicted in **Figure 1**, which evolved over many years in the pharmaceutical industry. Contemporary natural product screening programs are highly automated and organized "discovery engines" in which a large number of samples are processed on a continuous basis, by humans and a wide array of laboratory robots and instruments. Under most circumstances, natural products screening represents one part of a larger discovery effort (*e.g,* combinatorial synthesis, licensed compounds, in-house compound libraries all of which are tightly integrated within a given organization to feed samples into a common pipeline).

Screening programs use continuous feedback loops to learn from past experience. Throughout each program there are numerous decision points, each designed to exclude samples from further testing based on a failure to meet one or more predetermined criteria that are rooted in organizational knowledge. When a successful outcome is achieved (*e.g.* a new chemical entity with desirable drug-like properties), the entire chain of events leading to that discovery must be recovered to determine among other things, inventorship, regulatory compliance, and any obligations that might be due to third parties. Tracking each sample through the screening process is essential and links between the data and the physical sample must be accurately maintained throughout the process. So too, must be the transfer of materials and data from one individual to another. To accomplish this, most companies have developed sophisticated multi-user laboratory information management systems (LIMS) along with policies for their use. Unique identifiers[2] are used in many different ways in such systems and each is typically controlled by an internal central authority to ensure that access to current and legacy data and associated materials is maintained over long periods of time. LIMS are carefully curated and mapping of different identifiers is carefully maintained because the consequences for failure could be costly or disastrous to the organization. These systems are, however, invariably closed to the outside world for sound business reasons. Nonetheless, these LIMS are among the most complex examples of how identifiers are used to track genetic resources and the associated data, information, and knowledge through the entire research and development process; from the point of entry to the final decision point regarding that outcome and utility of that genetic resource for the intended purpose.

Much can be gained by understanding the various entry points for genetic resources into such programs, the types of samples that are created during the screening process, the types of samples that are transient, the types that are long-lived and potentially reused, and how identifiers are created and used. This can also provide some insight into what types of objects should be assigned PIDs by genetic resource providers, prior to transfer to users.

In **Figure 1.**, there are multiple entry and exit points for genetic resources. For the purposes of this discussion, we recognize a key distinction between viable genetic resources, non-viable genetic resources, and derivatives of genetic resources. This distinction has important ramifications on the genetic resource provider-user relationship. Viable genetic resources include any materials that may be preserved and repropagated by the user or a third party at some point in the future. Examples of such resources include but are not limited to purified microbial cultures. Non-viable genetic resources include whole or parts of plants or animals that cannot be propagated by the user. Purified nucleic acids represent an intermediate genetic resource. While nucelic acids cannot be used to recreate the source organism at present, it can be replicated in part, *in vivo* or *in vitro* to achieve desired goals and used in a variety of ways that are distinct from what is found in nature.

It should be stressed that under most circumstances it is essential that users be permitted to preserve, retain and repropagate genetic resources (as well as data derived from that resource) as a condition of use. Failure to recognize this need will disqualify a genetic resource provider's materials from use in such programs. Providers should clearly understood that international patent law requires the deposit of microorganisms, in viable form, in an officially recognized Biological Resource Center in support of any patent application arising from the use of that organism for purposes of enablement of claims. Once a patent issues, that material must be made irrevocably available to skilled artisans

[2] In LIMS, identifiers used for internal purposes need only be locally unique, as such identifiers are not intended for use in other systems. There are numerous examples of locally unique identifiers that appear in the STM and patent literature. These identifiers represent a chronic source of ambiguity, especially when legacy data is involved, as there is no guarantee that such identifiers are unique, persistent or will resolve to a physical or abstract resource. While local unique identifiers can be recast as globally unique identifiers, the conversion and validation process is time consuming and expensive and should be done only in those cases where the organization need justifies the cost and owner of the objects bearing those identifiers has a plan for curation and maintenance of that can guarantee persistence of the resolved objects.

to substantiate claims. Failure to make this deposit prior to filing a patent application will invalidate any claims. Therefore, it is useful to anticipate these needs in any agreements, and those agreements (PIC, MTA, CoO) should recognize these needs. PICs, MTAs, and CoO should permanently bound to the records of each genetic resource covered. Linkage between the PIC, MTA and CoO should also be considered essential.

Viable genetic resources can also be used as biocatalysts; to enzymatically modify synthetic, semi-synthetic and natural products in such a manner as to improve their desirable properties or mitigate undesirable properties. Viable genetic resources can be obtained specifically for this purpose or tested separately for biocatalytic properties in screening efforts. Once such strains are discovered, they are often added to libraries for use in the future. Information about biocatalytic properties is typically bound to the genetic resource when initially retained for future use as part of a biocatalytic tool-kit.

Non-viable resources are typically acquired in a form that is ready for chemical extraction prior to screening. Historically, samples were collected in sufficient quantity to permit multiple solvent extracts to be prepared for screening, rescreening, isolation and if possible, structure identification upon initial testing. This approach has been modified recently and now some providers are offering solvent extracts of their genetic resources for direct screening by users, pre-formatted in a manner to permit direct entry into automated screening programs (*e.g.*, dispensed into 96 or 384 well micro-well plates). It should be understood by providers that parts of whole organisms provided in non-viable form can be used for recovery of DNA by users to screen for genes of interest. Providers should contemplate such eventualities during contract negotiations and act accordingly, with the full understanding that those genes discovered by this route may be found in many other species as well.

Non-viable genetic resources may also include various environmental substrates from which microorganisms can be selectively isolated. It is not uncommon for a single sample of soil, sediment, leaf litter, or other material to yield tens to hundreds of bacteria and fungi, each of which can then enter the screening process as a wild-type isolate.

We consider derived genetic resources to be the full complement of products of gene expression that could trigger any detection assay. This includes proteins, lipids, carbohydrates, organic acids, or complex primary or secondary metabolites that might be discovered in any screening assay. Each such product is typically tracked in a LIMS system and, if sufficient material is available for rescreening, samples (or purified compounds derived from those samples) may be retained and rescreened in the future.

In the model presented in **Figure 1.**, identifiers are assigned to each microorganism that is selected for testing. Typically, this would be done by the laboratory responsible for the microbial isolation work. Those identifiers would point to strain specific information including source material and possible pointers to other identifiers (*e.g.* from a BRC, culture collection, or externally acquired collection) and any existing collecting agreement. When passed to the fermentation microbiologist, each organism is assigned a second identifier (often referred to as a batch number) that will bind all the screening results to the organism, the original internal source and any external provider (including any agreement). Each organism would be tested under a limited number of predefined growth conditions (typically 3 - 4) and each of those resulting samples (fermentation broths) would be subjected to a set of predefined solvent extractions (3-4). Thus, for each culture entering into the screens, 9-16 extracts would be tested in each of the primary assays. Typically, large scale screening programs would run 20-24 primary assays in parallel. Thus at this stage, a single culture could result in 160-288 specific samples for which data would become available. On average, screening programs typically examine 200-1000 isolates per week, so approximately 32-288,000 new samples would be generated weekly.

Samples that exceed certain thresholds in primary screens are then tested in more specific secondary assays to determine if the observed results are likely to be real. As before, additional identifiers need to be assigned prior to texting. Samples may also be sent for testing in dereplication systems, which results the assignment of yet more identifiers. If results at these stages of the screening process appear to be promising, then chemical isolation of bioactive compounds begins. As the chemist begins the fractionation process, additional samples are generated for testing in primary and secondary assays. If insufficient material is on hand to support isolation, a request for a additional fermentation broth is made, typically in larger volume. The microbiologist responsible for that task then prepares a second batch (typically) for which a new batch number is assigned and attached to the identifier chain indicating that the process of reproducing results has begun. Oftentimes, after the first two or three iterations through this "loop", strain improvement will also begin with the goal of increasing the yield of the compound of interest. This results in the assignment of many additional batch and sample numbers.

Should the chemical structure of a bioactive lead prove to be novel (or the utility of a known compound be novel), steps are typically taken to obtain patent protection. This results in the deposit of the producing strain in a recognized patent repository and securing BRC accession number for the organism, which appears in the patent application along with a description of the organism. Any nucleic acid or protein sequence data that is relevant to the invention or to assist in identifying the deposited strain may also appear in the patent application. While these public identifiers become part of the internal discovery record, the complete information trail remains hidden from view by outsiders.

As noted above, some genetic resources are frequently preserved and reused. Typically, large-scale screening programs have internal culture collections in which viable samples of microorganisms (wild-type, reference strains and various clone libraries, genomic DNA) and cell-lines are preserved for future use. These collections typically use their own identifier schemes that become integrated into the corporate LIMS. Derived genetic resources (from both viable and non-viable) may also be stored for future use (*e.g.*, compound and extract libraries) and are typically identified by compound numbers, which are used in subsequent screening. Finally, in some programs, additional identifiers are assigned when compounds advance from the unregulated research phase (discovery through initial scale-up, preliminary toxicity and pharmacology studies) to the more formal and tightly regulated preclinical and clinical stage of research. As each large-scale LIMS tends to be unique and designed to meet organizational needs and culture, the challenge becomes how to devise a CBD ABS tracking system that is interoperable with existing LIMS and other systems to ensure that appropriate information is available on a need-to-know basis.

Examples of numbering schemes that the authors have encountered in LIMS used in large-scale screening programs are presented in **Table 1**. Typically, these identifier schemes have varying amounts of imbedded information that may be useful to the issuing organization at the time of creation and may be subject to change to meet organizational needs. In each case, a designated authority within organization controls issuance of an identifier and each identifier is tracked by those authorities to insure uniqueness within the respective systems.

Table 1. Examples of numbering schemes used in LIMS

	Explanation of imbedded information
Sample Identifier 0534-605F	The first two digits indicate the year of isolation of a given organism, the second two the week of isolation, the next three digits indicate the screening assay in which a positive result was first observed and the final digit includes information about the organisms type. An identifier of this type is problematic as it mixes two types of information and does not support instances in which multiple bioactivities can be assigned to the same organism. If tied to a calendar year, it can only accommodate samples collected in a given century.
Sample Identifier B325,797-001-001	The first character was assigned to a block of identifiers reserved for use within a given program or subprogram, indicating responsible units within an organization. The first three digits indicate a subprogram tied to the personnel, biological resource types and sample providers. The next three digits indicate a confirmed bioactivity that had met multiple criteria to warrant assignment to a chemist for isolation and characterization. The next three digits represent unique batch of material prepared for screening and/or isolation, and the trailing digits represent unique samples that are generated by various individuals for testing. This is also an example of a compound identifier with embedded intelligence. The first eight characters in the string are equivalent to a genetic resource identifier and the remaining eight characters are sample identifiers that are tied to that genetic resource. However, this same number scheme is applicable to synthetic, semi-synthetic and purified, naturally occuring compounds, the latter being assigned an additional unique identifier once the purified compound is obtained. The only distinction being the first three digits.
Strain identifier XX-9999a	The first two characters identify a collection and subcollection arranged on the basis of the broad grouping of organism types. The designations oftentimes appear somewhat connected with more widely accepted classifications as some organisms (e.g. Actinobacteria and yeasts) are split away from phylogentically related taxa for internal purposes. The trailing letter indicates a passage series, and is incremented when stocks preserved stocks are replenished by making successive "second series" from the original preserved genetic resource. Such identifiers typically map to portions of the sample identifiers described above and to accession numbers from various BRCs, external culture collections, and other providers from whom the original material was obtained.

Discussion

At this point in time, it is unlikely that the large-scale bioprospecting efforts that were central to the drug discovery programs in many of the large North American, European, and Japanese pharmaceutical companies during the last 20 years will return in the same form. Most companies have abandoned that route to drug discovery because of the high cost and complexity as compared to alternative strategies (*e.g.*, combinatorial chemistry, genomics, siRNA) and the presumed failure of that approach to yield new chemical entities that led to blockbuster drugs. A similar fate has been met by smaller research organizations that attempted to pursue this line of discovery. While one can debate that vast riches lie in wait, until some unspecified breakthrough occurs by this route to drug-discovery it is likely to be deemphasized in favor of more promising approaches.

On the other hand, it is likely that genetic resources will receive greater interest as a source of many new commodity products and processes, ranging from biofuels to genes and gene products for use in a wide variety of new industrial process. Bioprospecting will also shift from the domain of field biologists to bioinformatists and computational biologists who will mine ever increasing amounts of genomic and related data in public and private repositories for the bits and pieces that can be incorporated into more readily controlled expression systems to achieve the desired results. In many cases, these will result in novel chimeras having little resemblance to their progenitors.

Regardless of the route to discovery or the type of product, we believe that most genetic resource providers and users are making good faith efforts to find an efficient way to reach ABS agreements that preserve their own interests, without jeopardizing those of the other parties. There are thus players who are genuinely expecting a workable IR that takes into account the needs and expectation of all involved parties. The challenge for the negotiation process is to gain wisdom and insight from the past to create workable solutions that are open, interoperable and extensible so that the technical changes on the near to distant horizon are accommodated in a manner that is beneficial to all parties.

Since the CBD was drafted, the life sciences have undergone radical change because of technological advancements in a wide range of fields. Genetic resources are routinely rendered in more informational forms such as extracted DNA as well as gene and genome sequences held in public and private databases. As a result, genetic resources are much more mobile and therefore much more accessible to a global audience. The next section examines recent advances in DNA sequencing technologies that enable the identification of biological organisms at the genetic level.

References

Baltz, R.H. (2007) Antimicrobials from Actinomycetes: Back to the future. *Microbe*, 2(3), 125-131.

Beehler, B.M. (2007) The Foja Mountains of Indonesia: Exploring the Lost World. *ActionBioscience.org*, January 2007. Available at: http://www.actionbioscience.org/biodiversity/beehler.html

Burton, G. and Phillips, B. (2005) *Developing a system of Virtual Certificates of Origin and Provenance*. Paper at the International Expert Workshop on Access to Genetic Resources and Benefit-sharing. September, 2005. Cape Town, South Africa.

Channon, K., Bromley, E. H. and Woolfson, D. N.. (2008) Synthetic biology through biomolecular design and engineering. *Curr.Opin.Struct.Biol*. 18:491-498. doi:10.1016/j.sbi.2008.06.006

Christoffersen, L.P. and Mathur, E.J. (2005) Bioprospecting Ethics & Benefits: A model for effective benefit-sharing. *Industrial Biotechnology*, 1(4): 251-258. Available at: PDF

Convention on Biological Diversity (2002) *Bonn Guidelines on Access to Genetic Resources and Fair and Equitable Sharing of the Benefits Arising out of their Utilisation. Decision VI/24.* Available at: PDF

Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R. andWittwer, P.. (2005) Metagenomic gene discovery: past, present and future. *Trends Biotechnol.* 23:321-329. doi:10.1016/j.tibtech.2005.04.001

Dalton, R. (2004) Bioprospectors hunt for fair share of profits. *Nature*, 427(6975): 576. doi:10.1038/427576a

Daniel, R. (2004) The soil metagenome--a rich resource for the discovery of novel natural products. *Curr.Opin.Biotechnol*. 15:199-204. doi:10.1016/j.copbio.2004.04.005

Feit, U., Lobin, W. and Driesch, M. (2005) *Access and Benefit-Sharing of Genetic Resources: Access and Benefit-Sharing of Genetic Resources: Ways and means for facilitating biodiversity research and conservation while safeguarding ABS provisions. Report of an international workshop.* Bonn, Germany: German Federal Agency for Nature Conservation. Available at: PDF

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C. et.al., (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl.Environ.Microbiol.* 73:7059-7066. doi:10.1128/AEM.00358-07

Gabor, E. M., de Vries, E. J. and Janssen, D. B. (2004) Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases. *Environ.Microbiol*. 6:948-958. doi: 10.1111/j.1462-2920.2004.00643.x

Keasling, J. D. (2008). Synthetic biology for synthetic chemistry. *ACS Chem.Biol*. 3:64-76. doi: 10.1021/cb7002434

Klinkenborg, V. (2008) Watching the Numbers and Charting the Losses - of Species. *The New York Times*. October 15, 2008.

Kuo, A. and Garrity, G.M. (2002) Exploiting Microbial Diversity. In Staley, J. and Reysenbach, A.L. (eds) *Biodiversity of Microbial Life-Foundation of Earth's Biosphere.* New York: John Wiley & Sons, Inc.

Laird, S.A. and Wynberg, R. (2008a) Access and Benefit-Sharing in practice: Trends in partnerships across sectors, Montreal, Canada: Secretariat of the CBD. Available at: PDF

Laird, S.A. and Wynberg, R. (2008b) Study on access and benefit sharing arrangements in specific sectors. Prepared for the Ad Hoc Open-Ended Working Group on Access and Benefit-Sharing, Sixth meeting, Geneva: Switzerland, 21-25 January. Available at: PDF

Laird, S.A. and Wynberg, R. (2008c) Bioprospecting: securing a piece of the pie. World Conservation IUCN, January 2008: 27-28. Available at: PDF

Lämmle, K., Zipper, H., Breuer, M., Hauer, B., Buta, C., Brunner, H., and Rupp, S. (2007) Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *J.Biotechnol.* 127:575-592.

Leonard, E., Nielsen, D. Solomon, K. and Prather, K. J. (2008). Engineering microbes with synthetic biology frameworks. *Trends Biotechnol.* 26:674-681. doi:10.1016/j.tibtech.2008.08.003

Li, G., Wang, K. and Liu, Y. H. (2008) Molecular cloning and characterization of a novel pyrethroid-hydrolyzing esterase originating from the Metagenome. *Microb.Cell Fact.* 7:38. doi:10.1186/1475-2859-7-38

Maron, P. A., Ranjard, L., Mougel, C., and Lemanceau. P. (2007) Metaproteomics: a new approach for studying functional microbial ecology. *Microb.Ecol.* 53:486-493. doi: 10.1007/s00248-006-9196-8

McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat.Methods* 4:63-72. doi:10.1038/nmeth976

Meyer, A., Pellaux, R. and Panke, S. (2007) Bioengineering novel in vitro metabolic pathways using synthetic biology. *Curr.Opin.Microbiol.* 10:246-253. doi:10.1016/j.mib.2007.05.009

Muller, M. and Lapeña, I. (2007) A Proposal on international audits to track and monitor flows of genetic resources and verify compliance with ABS Agreements. In *A Moving*

Target:Genetic Resources and Options for Tracking and Monitoring their International Flows. ABS Series. Gland, Switzerland: IUCN, p. 111-123. Available at PDF

Newman, D.J. and Cragg, G.M. (2007. Natural products as sources of new drugs over the last 25 years. *Journal of Natural Products*, 70(3), 461-77. doi: 10.1021/np068054v

Newman, D.J., Cragg, G.M. and Snader, K.M. (2003) Natural Products as Sources of New Drugs over the Period 1981−2002. *Journal of Natural Products*, 66 (7): 1022-1037.doi: 10.1021/np030096l

Noirel, J., S. , Ow, Y., Sanguinetti, G., Jaramillo, A. and Wright, P. C. (2008) Automated extraction of meaningful pathways from quantitative proteomics data. *Brief.Funct.Genomic.Proteomic*. 7:136-146. doi:10.1093/bfgp/eln011

NZPA (2008) Scientist finds new monkey. *Herald on Sunday*. Retrieved December 01, 2008 from: http://www.nzherald.co.nz/science/news/article.cfm?c_id=82&objectid=10490946

O'Malley, M. A., Powell, A., Davies, J.F. and Calvert, J. (2008) Knowledge-making distinctions in synthetic biology. *Bioessays* 30:57-65. doi: 10.1002/bies.20664

Parry, B. (2004. *Trading the genome: Investigating the Commodification of Bio-information*. New York: Columbia University Press.

Peccoud, J., Blauvelt, M.F, Cai, Y., Cooper, K.L., Crasta, O., et.al., (2008) Targeted development of registries of biological parts. *PLoS.ONE*. 3:e2671. doi:10.1371/journal.pone.0002671

Petersen, F. and Kuhn, T. (2008) How to link bioprospecting with sustainable capacity building. *Business.2010*, 3(1): 14. Available at: PDF

Pham, V. D., Palden, T. and DeLong, E. F. (2007) Large-scale screens of metagenomic libraries. *J.Vis.Exp*. 2007:1-2. doi: 10.3791/201

Pisupati, B. (2007) *UNU-IAS Pocket Guide: Access to Genetic Resources, Benefit Sharing and Bioprospecting*, Yokohama: UNU-IAS. Available at: PDF

Rai, A. and Boyle, J. (2007) Synthetic biology: caught between property rights, the public domain, and the commons. *PLoS Biology* 3(5):389-393. doi:10.1371/journal.pbio.0050058

Riesenfeld, C. S., Schloss, P. D. and Handelsman, J. (2004) Metagenomics: genomic analysis of microbial communities. *Annu.Rev.Gene*t. 38:525-552. doi:10.1146/annurev.genet.38.072902.091216

Smagadi, A. (2005) National measures on access to genetic resources and benefit sharing-The case of the Philippines. *LEAD: Law Environment and Development Journal*, 1(1). Available at: PDF

Steele, H. L., Jaeger, K. E., Daniel, R. and Streit, W. R. (2009) Advances in recovery of novel biocatalysts from metagenomes. *J.Mol.Microbiol.Biotechnol*. 16:25-37. doi: 10.1159/000142892

Swiss Academy of Sciences (no date). A short introduction to the rules governing access to genetic resources and the cases in which these rules apply. Retrieved October 2008, from http://abs.scnat.ch/basics/index.php

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., and Glockner, F.O. (2004.) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ.Microbiol*,. 6:938-947. doi: 10.1111/j.1462-2920.2004.00624.x

Tobin, B. (2005) The power of collecting. *Nature*, 435(7044), 887. doi: 10.1038/435887a

Tobin, B., Burton, G. and Fernandez-Ugalde, J.C. (2008) *Certificates of Clarity or Confusion:The search for a practical, feasible and cost effective system for certifying compliance with PIC and MAT.* Japan: UNU-IAS. Available at: http://www.ias.unu.edu/sub_page.aspx?catID=111&ddlID=682

Tulp M, Bohlin L. (2002) Functional versus chemical diversity: is biodiversity important for drug discovery? *Trends Pharmacol Sci*., 23(5): 225-31. doi: 10.1016/S0165-6147(02)02007-2

WCS (2007) Discovery on uncharted lands. Retrieved November 01, 2008 from http://www.wcs.org/353624/wcs_kabogo

Wilkinson, B. and Micklefield, J. (2007) Mining and engineering natural-product biosynthetic pathways. *Nat.Chem.Biol*. 3:379-386. doi:10.1038/nchembio.2007.7

Williams, C., Davis, K., and Cheyne, P. (2006) *CBD for Botanists: An introduction to the Convention on Biological Diversity for people working with botanical collections*, Kew, London: Royal Botanic Gardens. Available at: http://www.kew.org/data/cbdbotanists.html

Wilmes, P. and Bond, P. L. **(**2006) Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol*, 14:92-97. doi:10.1016/j.tim.2005.12.006

Wilson, Z. and Brimble, M. (2008) Instant insight: Life at the extremes. *Chemical Biology*, (12) 11 November 2008. Retrieved December 15, 2008 from http://www.rsc.org/Publishing/Journals/cb/Volume/2008/12/Extremophiles.asp

ADVANCES IN GENETIC IDENTIFICATION

Before any species of plant, animal or microbe can be protected or become part of a sustainable development effort, it must first be identified and given at least a trivial name so as to distinguish it from all others. Careful and detailed taxonomy (the science dealing with the description, identification, naming, and classification of organisms) has an integral role in achieving goals of the CBD. Given that we now realize that there can be a convergence in phenotype among unrelated species, taxonomies based solely on morphological analyses can result in misclassification (Lorenz *et al.,* 2005). Morphology is somewhat useful for work with higher eukaryotes that undergo discernable differentiation into well-defined and observable body structure. However, in the case of bacteria and archaea, the extent of observable differentiation is much less and the extent of homoplasy is much greater.

The development of molecular technologies that allow for identification of organisms at a genetic level opened new possibilities for taxonomic research (Lorenz *et al.,* 2005). Genetic identification of an organism is basically a comparative process, which in theory is relatively straightforward. However, in practice, the process is more complex and dynamic.

In principle, to identify an unknown organism, appropriate sequences from the unknown are compared to homologous sequences from a reference set of known organisms (oftentimes taxonomic type material and other well characterized reference organisms) on a pair-wise basis. Similarity between the sequences above a predetermined (and often arbitrary) threshold results in a presumptive identification. When genomic sequences of various organisms are examined, related individuals have genetic material that is identical for some regions and dissimilar for others. Unrelated individuals have significant differences in many of the sequences being evaluated. The challenge is that the regions of sequence similarity and dissimilarity within different taxonomic groups tend to vary, oftentimes to a considerable extent. This is particularly true for groups of organisms that are not well characterized or for which only a few known representatives exist and are available for study. Often times, earlier identifications based on morphological or other phenotypic characters may have resulted in classification errors that further confound attempts to apply genetic techniques to identification. Developing a well-curated database of key sequences that are unique to and characteristic of a series of known organisms is critical to this type of analysis.

One challenging issue is the genomic diversity that exists within a given species. For example, within the human population more than three million locations along the human genome have thus far been identified in which one base differs from one person or population to the next (Pennisi, 2007). These are known as single-nucleotide polymorphisms (SNPs) and SNPs are being charted in what is known as the HapMap. There is also the challenge presented by those areas along the genome which are common among different species, such as the 600 protein domains that are found both in *E. coli* and in humans (InParanoid, 2008).

Since the first genome of a bacteriophage was sequenced in 1977 (Sanger *et al.,* 1977), rapid advancements in molecular technologies have had a dramatic impact on the speed and accuracy of DNA sequencing. In 1994, an international collaboration was established to sequence the genome of the yeast *Saccharomyces cerevisiae* (Levy, 1994). However, the first genome to be completely sequenced from a free-living organism was from the bacterium *Haemophilus influenzae* in 1995 by Fleischmann *et al.* at The Institute for Genomic Research, now know as the J. Craig Venter Institute (JCVI) using the whole genome shotgun (WGS) method. Data from this project, which included 1,830,137·bp of DNA and comprising 1743 predicted genes, laid out, for the first time, the full genetic complement of a bacterium (Fleischmann *et al.*, 1995, 2003).

Compared to bacteria and yeast, the genomes of most plants and animals are considerably larger. For example, the human genome is a thousand times the size of an average bacterial genome. The sequencing of the first rough draft of a human genome through the international Human Genome Project (HGP) (Venter *et al.*, 2001; Lander *et al.*, 2001) took several years to accomplish. When it first began in the mid 1980s it was quickly recognized that it would take more than a thousand years to accomplish this task using existing sequencing technology, bioinformatics tools and computer hardware. In light of this, heavy investments were made in improving the speed and accuracy of sequencing technologies. In fact, most of the $200,000,000 per year spent on this 20-year project was targeted towards improving the necessary technologies (Venter *et al.*, 2001; von Bubnoff, 2008). As a result, continual improvement in the methodology of DNA sequencing, biochemistry, bioinformatics and other related fields have reduced the time required and the cost per base pair while also greatly improving the accuracy and quality of the final product (Fraser *et al.* 2002; Shendure, 2008).

The Genomes on Line Database (GOLD) was created in 1997 to continuously monitor genome sequencing projects worldwide and provide the community with a unique centralized resource integrating information related to *Archaea*, *Bacteria*, *Eukaryia* and metagenome sequencing projects (Liolios *et al.*, 2008). As of January 10, 2009 there are 921 published genome projects currently registered on the GOLD: 55 of these are archaeal; 766 are bacterial; and 100 are eukaryotic. These sequences are deposited into the International Nucleotide Sequence Database Collaboration (INSDC), as is required by various public and quasi public funding agencies, non-governmental organizations and private foundations that underwrite these initiatives. The INSDC is comprised of three major centers dedicated to this task: the DDBJ , EMBL and GenBank (Liolios *et al.*, 2008).

The sequencing method developed by Sanger and his colleagues (1977), also referred to as the classical chain-termination method, has been the mainstay of sequencing technology for almost thirty years. The Sanger sequencing method can achieve read-lengths of up to 1000 nucleotides (nts) at a cost of approximately fifty cents ($0.50 U.S.) per kilobase (Shendure, 2008). Continued refinements in capillary electrophoresis systems combined with increased laboratory automation, process parallelization, software improvements, and the corresponding development of databases of complete reference genome sequences has steadily driven down the costs of sequencing over the years, as is shown in **Figure 2**. (Blow, 2008; Holt, 2008; Schuster, 2008; Service, 2006).

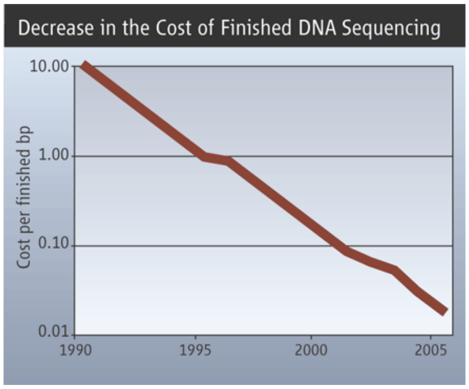

Figure 2 Free fall. As with computer technology, the plunging cost of DNA sequencing has opened new applications in science and medicine. Source: Service (2006, pg 1544) in *Science*. Reprinted with permission from AAAS

Although the cost continues to drop, there are still physical limitations to the Sanger dideoxy method, and in practice completely sequencing genomes using this method still takes a considerable investment of time, effort, and money. There have been a number of global initiatives to stimulate the development of what has become commonly referred to as second or next-generation sequencing (NGS) technologies. These include The Yanhuang Project in China, The 1,000 Genome Project, the Genomic X Prize and the 500-Euro-human Genome:

> The Yanhuang Project, which will sequence drafts of genomes of 100 Chinese individuals over 3 years, was announced by the Beijing Genomics Institute (BGI) in January 2008 (Qiu, 2008).
>
> In January 2008 the 1,000 Genome Project was announced by a new international research consortium (National Institutes of Health (NIH) in the United States, the Wellcome Trust Sanger Institute (Sanger) in the United Kingdom; and BGI in China). The project is expected to cost between $30 - $50 million, and its overall goal is to uncover more detailed genetic factors involved in human health and disease. The consortium will sequence genomes from at least 1,000 volunteers worldwide to ensure representation of African, Asian and European populations; thereby producing a catalog of human variation down to variants that occur at 1% frequency or less over the genome, and 0.5–0.1% in genes. The National Human Genome Research Institute (NHGRI) of the NIH in the United States will support and fund three of the large genome centers which will primarily be responsible for producing sequence data for the project (Siva, 2008; Hayden, 2008; von Bubnoff, 2008).
>
> The J. Craig Venter Science Foundation (JCVI) joined forces with the X PRIZE Foundation (X PRIZE) in 2006 to create the $10 million Archon X PRIZE for Genomics competition. Like the original human genome project the purpose of this competition is to stimulate development of new technology that will reduce the time and cost of sequencing. The Prize winner will be the first group that sequences 100 human genomes in 10 days or less, with an error rate below 10^{-5}%, coverage at > 98% of the genome, and a recurring cost of no more than $10,000 per genome (Archon, 2008).

> In Europe, the commercial firm GATC has a vision of "500-Euro-human genomes". It is offering "its DNA analysis expertise and sequencing capacity to strategic partners from academia and industry, and to private pioneers that are interested in sequencing up to 100 human genomes by the end of 2010" (GATC, 2008).

NGS technology aim to address a number of key issues (Chan, 2005, Schuster, 2008):

- Cost reduction to deliver on the promise of $1,000 personal genomics
- Improvement of read length, throughput and data output quality
- Resolution of the bioinformatic bottleneck arising from the exponential increase of data produced by these methods.

Ideally, data analysis also includes genome annotation: the process in which the locations and functioning of genes and all of the coding regions in a genome are identified. The results are then reported in one of the gene banks such as EMBL or GenBank. With the current short reads, unannotated sequence data is being reported as well. Furthermore, genomes are not all annotated to the same standards and the quality of annotation of the genes can vary greatly as well (Ussery and Hallin, 2004).

What follows is a "snapshot" of the current state of the field. It should be understood that many new NGS methods are continually being developed, and current methods are continuously refined to remain competitive on throughput, accuracy, read-length and sequencing cost/kilobase. The first part of this section covers the currently used NGS technologies which include sequencing on a microchip: 454, Solexa and SOLiD™ . The next part looks at those systems that have just recently become available on the market, *i.e.*, the near-future NGS methods, the Polonator and the tSMS™ and the final part reviews those NGS technologies that are still under development.

Current NGS technologies

Ultrafast DNA sequencing on a microchip by a hybrid separation mechanism

Among the technologies being developed to reduce sequencing costs, microchip electrophoresis is one that is able to produce long reads most suitable for the *de novo* sequencing and assembly of large and complex genomes. Microchip systems promise to reduce sequencing costs dramatically by increasing throughput, reducing reagent consumption, and integrating the many steps of a sequencing pipeline onto a single platform. Fredlake *et al.* (2008) reported "sequencing up to 600 bases in just 6.5 min by microchip electrophoresis with a unique polymer matrix/adsorbed polymer wall coating combination". These results represent a two-thirds reduction in sequencing time over any comparable chip sequencing result. These authors explain their ultrafast long reads on chips as follows:

> the combined polymer system engenders a recently discovered "hybrid" mechanism of DNA electromigration, in which DNA molecules alternate rapidly between reptating through the intact polymer network and disrupting network entanglements to drag polymers through the solution,

> similar to dsDNA dynamics we observe in single-molecule DNA imaging studies. Most importantly, these results reveal the surprisingly powerful ability of microchip electrophoresis to provide ultrafast Sanger sequencing, which will translate to increased system throughput and reduced costs (Fredlake *et al.*, 2008, p. 476).

Most of the new technologies are continually seeking to miniaturize, multiplex, and automate the sequencing reaction even further (Blow, 2008; Gupta, 2007; Service, 2008). Additionally, capillary sequencing used in the Sanger method is no longer the technology of choice for most of the NGS ultra-high-throughput applications.

Flow sequencing

A new generation of instruments utilizes primed synthesis in flow cells to simultaneously obtain the sequence of millions of different DNA templates; this is an approach that has changed the field (Hall, 2007; Holt *et al.*, 2008). Holt and Jones (2008) observed that "if the hallmark of the past paradigm was electrophoretic separation of terminated DNA chains, then the hallmark of the new paradigm is flow sequencing, with stepwise determination of DNA sequence by iterative cycles of nucleotide extensions done in parallel on massive numbers of clonally amplified template molecules" with the end result of DNA being sequenced much faster and cheaper.

Flow sequencing, also known as sequencing by synthesis (SBS) on a solid surface, tracks nucleotides as they are added to a growing DNA strand (Blow, 2008; Käller *et al.,* 2007). SBS is used by the following ultra-high-throughput sequencing systems which have become commercially available in the past two years:

- Genome Sequencer 20/FLX/Titanium (commercialized by 454/Roche);
- Solexa1G' (later named 'Genome Analyzer' and commercialized by Illiumina/Solexa);
- SOLiD™ system (commercialized by Applied Biosystems).

The ultra-high-throughput sequencing systems used most extensively today are the three systems mentioned above. These developments have increased the sequencing speed while significantly reducing the cost of sequencing (Gupta, 2008; Davies, 2008). With the current platforms, the much higher throughput means greater coverage but at the cost of much shorter read-lengths. Hence, a whole new set of issues must be addressed with the new technologies which include higher error-rates, the analysis of massive data sets that are output by these systems and short read-lengths that complicate assembly, especially of eukaryotic genomes in which large number of repeat sequences occur.

The NGS methods to date generate read lengths ranging from 35 nucleotides (nts) with the Solexa to more than 500 nts using the 454-Titanium, which are significantly shorter than the 800-1000 nts reads that are typically achieved by the Sanger method (Smith, 2008; Ten Bosch, 2008; Gupta, 2008). One reason that next-generation technology is being eyed for other biological applications is its increase in throughput (Smith, 2008).

The first step in most sequencing processes is DNA amplification because it is extremely difficult to measure biochemical processes at the level of only a few molecules (Hall, 2007). In the Sanger method the DNA is usually cloned into bacterial plasmids. However, producing clones has its own set of problems. Two sequencing methods—454 array-based pyrosequencing (Margulies *et al.*, 2005) which is currently in wide use, and polony sequencing (Shendure *et al.*, 2005)—have developed high-throughput strategies for *in vitro* amplification that are very inexpensive and circumvent the inherent biases of *in vivo* methods used in the Sanger method (Ronaghi, 2001; Elahi *et al.*, 2004; Hall, 2007; Holt, 2008) The 454 Flex/Titanium pyrosequencing platform by 454 Life Sciences (now owned by Roche Diagnostics), can generate 200 million nucleotides data in a 7 hour run with a single machine.

Schuster (2008) notes "both approaches use a strategy that greatly reduces the necessary reaction volume while dramatically extending the number of sequencing reactions. The strategy entails arraying several hundred thousand sequencing templates in either picotiter plates or agarose thin layers, so that these sequences could be analyzed in parallel—a huge increase compared to the maximum of 96 sequencing templates on a contemporary Sanger capillary sequencer".

•The first of the massively parallel methods to become commercially available was from 454 Life Sciences in 2005. The 454 Titanium instrument carries out pyrosequencing (Margulies *et al.,* 2005) reactions in parallel. The pyrosequencing process is outlined by Service:

> This process first fragment a genome into stretches 300 to 500 base pairs long, denatures the DNA, discards one strand, and links the other to a functional group that is tethered to a plastic bead--at a concentration such that each bead gets just one strand. These fragments are then replicated by the polymerase chain reaction (PCR) until the copies cover each bead. The beads are separated on a plate containing as many as 1.6 million wells and dosed with a series of sequencing reagents and nucleotides. Every time a nucleotide is added onto a growing DNA chain, the reaction triggers the release of a pyrophosphate group, which in turn prompts a firefly enzyme called luciferase in the well to give off a flash of light. By correlating the recorded flashes from each cell with the nucleotides present at the time, a computer tracks the systematic sequence growth of hundreds of thousands of DNA fragments simultaneously. This system allows shotgun sequencing of whole genomes without cloning in *E. coli* or any host cell (2007, p. 1545).

•The Illumina Genome Analyzer can produce 600 Mb sequence DNA per day. This sequencer achieves parallelization by the *in situ* amplification of DNA fragments immobilized onto the flow cell of the instrument at a concentration that promotes a dense array of non-overlapping fragment colonies. The sequencing differs from polony or 454 pyrosequencing as it amplifies the DNA on a solid surface followed by synthesis by incorporation of modified nucleotides linked to colored dyes. (Hall, 2007, Hutchinson, 2007, Holt 2008; Hillier *et al.*, 2008; Smith *et al.*, 2008). Nucleotides with four different colors and standard microarray optics are used to track the growth of strands complementary to those attached to the slide (Service, 2006). Currently, the second generation Solexa machines can produce read lengths of approximately 70 nts (Quail, *et al.*, 2008).

•The SOLiD platform uses sequencing by ligation, which produces DNA sequence by measuring the serial ligation of an oligonucleotide. All fluorescently labeled oligonucleotide probes are present simultaneously and compete for incorporation. After each ligation, the fluorescence signal is measured and then cleaved before another round of ligation takes place. A reset phase allows a reduction in noise—a capping step that prevents dephasing.

Near-future NGS methods

•The Polonator evolved from a collaborative effort between Dover Systems and the Church laboratory of Harvard Medical School. It is a completely open platform, combining a high performance instrument, with freely downloadable, open-source software and protocols, off-the-shelf reagents, and inexpensive flow cells. It is based on polony sequencing technology. The current short read-lengths of 26 nt however, is significantly limiting (Service, 2006).

As mentioned previously, the significantly shorter read lengths of current NGS methods is problematic, especially when compared to the much longer reads produced by the Sanger method. The short segments make it difficult to reassemble all the pieces into a continuous genome sequence (Chaisson *et al.*, 2004). This reassembly is further complicated by the fact that there are short repeats of a couple of hundred nucleotides that may be found thousands perhaps millions of times throughout the DNA strand. Another drawback is that these methods rely on PCR, which is expensive and can introduce copying errors. Greater experience with the new sequencing technologies may improve matters. Several groups are developing ways to sequence a single copy of a long DNA strand, thereby achieving longer reads and avoiding PCR (Service, 2006).

It is expected that what has become known as the next-next generation sequencing, also known as third generation sequencing systems will address many of these drawbacks since they are based on true single-molecule sequencing (SMS) which do not involve amplification steps (Gupta, 2008) in which an entire DNA strand can be sequenced. SMS has been realized in the laboratory through several approaches, such as scanning probe microscopy, exonuclease sequencing, and sequencing by synthesis (SBS), among others. The first commercial SMS instrument was launched in 2008 by Helicos Biosciences, (Gupta, 2008; Hall, 2007; Blow, 2008).

•The true single-molecule sequencing (tSMS™) of Helicos Biosciences (Cambridge, MA, USA) is a proprietary method in which a SBS approach for single molecules is implemented in a HeliScope single-molecule sequencer. It is the first true single molecule sequencing platform to become commercially available. In this system*,*

> the target DNA is used for the construction of a library of poly(dA)-tailed templates, which pair with millions of poly(dT)-oligonucleotides that are anchored to a glass cover-slip. The positions of each of these individual poly(dT) oligos – and hence those of the respective paired poly(dA)oligos – on the cover slip are determined by camera imaging. The sequence of each poly(dA)-tailed fragment is determined by adding nucleotides – each labeled with the same cyanine dye Cy5 (a non-radioactive fluorescent dye) –in a cyclic manner, one at a time. The incorporation of nucleotides to each poly(dT) – or, indeed, the lack of incorporation, depending upon complementarity – enables faithful copying of the paired poly(dA)-tailed

> templates for sequencing. The events of nucleotide incorporation are imaged with a camera and used to obtain 30-base-long sequences for each paired poly(dA)-tailed fragment. Error rates may be reduced by performing a so-called ‘twopass’ sequencing (Gupta, 2008, p. 604).

NGS methods under development

The methods presented here apply SMS by a wide variety of innovative approaches. Single-molecule sequencing should result in long, fast reads since it involves reading a single piece of DNA. This means that the problems inherent with the short reads produced by the current NGS platforms would be eliminated. However, there still remain the issues of error rates, including error rate determination, and the issue of analyzing the massive data output.

•*Fluorescence resonance energy transfer*

•The FRET-based SMS-SBS approach of VisiGen Biotechnologies (Houston, TX, USA) employs “a novel platform [that] uses single-molecule DNA detection, fluorescent molecule chemistry, computational biochemistry, and biomolecule engineering and purification” (Visigenbio, 2008). It uses fluorescence resonance energy transfer (FRET) for detection of incorporated nucleotides. In this method the terminal phosphate of a nucleotide is tagged with a fluorophore that is naturally released during nucleotide insertion into the growing DNA strand, thereby enabling a non-cyclical approach to DNA sequencing. The DNA sequence is read in real time by monitoring an engineered polymerase containing a donor fluorophore as it incorporates bases into a DNA strand. When a nucleotide, which has one of four differently colored acceptor flurophores attached to its gamma phosphate, is incorporated, the proximity of donor and acceptor fluorophores results in a FRET signal. The DNA molecule lights up, and the color indicates the base identity because the fluoropheres on the nucleotides are color coded. Each time a nucleotide is incorporated, the pyrophosphate containing the fluorophore is released so that the nascent strand synthesized is natural DNA, and no additional processing is needed before the next nucleotide can be incorporated (Gupta, 2008; Chan, 2006; Blow, 2008).

Since the tSMS technology of Helicos Biosciences requires the cyclic addition of reagents thereby increasing the time and cost of sequencing, one might consider that the FRET-based SMS–SBS approach is an improvement. VisiGen is planning to release its first sequencer in the market in 2009, and proposes a sequencing speed of one million bases per second, which would mean that an individual human genome might be fully sequenced within an hour. That is, at 1x coverage – but with the high error rates, one would presumably want much higher coverage rates. But nonetheless, the possibility of an overnight sequencing of a complete human genome is certainly tantalizing. Assembly and annotation would, however, remain a bottleneck.

•Pacific Biosciences (PacBio) announced a next-generation sequencing instrument that utilizes its single molecule, real time (SMRT) sequencing technology will be available in 2010 and will be capable of sequencing a diploid human genome at 1-fold coverage in about 4 minutes (von Bubnoff, 2008; Karow, 2008). This technology purportedly enables real time observation of natural DNA synthesis by a DNA polymerase and is made possible by the development and use of the SMRT chip (each containing many zero-mode waveguides (ZMWs) and phospholinked nucleotides (PacificBioscience, 2008; Levene, 2003).

•*Nanopore sequencing:*

Another method of single-molecule sequencing that is in the very early stages of development involves "reading" DNA as it is passed through a nanopore in a very thin membrane (Kasianowicz *et al.*, 1996; Storm *et al.*, 2005a; Storm *et al.*, 2005b). Nanopore-sequencing technologies do not involve an enzymatic extension reaction of any kind. These methods sequence DNA strands as they move through a tiny pore and read the bases either electronically or optically (Blow 2008; Service, 2006; Shendure, 2008; Gupta 2008). In theory, this method should have no limit on read length and, hence, if the technical hurdles are overcome, could revolutionize genome sequencing (Hall, 2007). Branton and colleagues (2008) believe that a parallelism of 100 nanopores is reasonable and that the resulting platform will be able to sequence a mammalian genome in twenty-four hours for less than one thousand dollars ($1000 U.S.; Blow, 2008).

One problem with the nanopore technology is that the DNA molecule might pass through the nanopore too quickly to enable the resolution of individual bases. There have been two possible solution to the speed problem offered: Hybridization-assisted nanopore sequencing (HANS) which is a proprietary technique developed by company NABsys in a joint venture with Brown University; and Design polymer-assisted nanopore sequencing which technology has been developed by LingVitae (Oslo,Norway) and involves the conversion of target DNA into a magnified form, the so-called 'design polymer' (Blow, 2008). Conceptually, another way to achieve this would be through immobilizing an enzyme (*e.g.*, an exonuclease or a poron) bound to one of the surfaces that could pull the DNA molecule through the pore at a steady rate.

•*Transmission electron microscopy for DNA sequencing*

In this SMS platform being developed by ZS Genetics (ZSG; North Reading, MA, USA) DNA sequences are read directly with the help of a specialized transmission electron microscope (TEM). This approach has been accepted as one of the competitors for the previously mentioned 'Archon X Prize'. The company claims that with its "sequencing technology, a sample is prepared once, a picture is taken with a transmission electron microscope and the sequence is read directly from that picture…[They expect] to be able to read over 1.7 billion base-pairs per day as a starting point, compared to 100 million for the leading Next-Gen competitor today" (ZSG, 2008). TEM technology entails

> the linearization of the target DNA molecule, followed by synthesis of a complimentary strand, whereby three of the four bases are labeled with heavy atoms, and the fourth base remains unlabeled. Given that atoms such as C, O, N, H and P present in DNA have low atomic number (Z = 1–15), natural DNA is transparent when viewed with TEM. However, bases labeled with heavier elements, with high Z values (*e.g.,* iodine with Z = 53; bromine with Z = 35), make the DNA heavier [electron dense]

> and, therefore, visible under TEM. Thus, when the resulting complementary strand is observed under TEM, the four bases can be discriminated by the size and intensity of dots representing the four bases (Gupta, 2008 p. 609).

•*Sequencing with nano-knife edge probes*

In its search for a new and improved DNA sequencing tool, Reveo developed the concept of the Omni Molecular Recognizer Application (OmniMoRA). This technology "is based on arrays of nano-knife edge probes that directly and non-destructively read the sequence of 3 billion base-pairs that comprise the human genome" (Reveo, 2008). Multiple nano-knife edge probes, each of which recognizes only one nucleotide, pass over a strand of DNA which is stretched and immobilized in a channel that is 10 micrometers wide. A unique voltage is applied to each nano-knife edge probe, and when the probe touches the corresponding nucleotide, electrons tunnel into the molecule, losing energy which is measured. When a nano-knife edge probe touches the wrong base no current is detected (Blow, 2008). Faris and Eakin (2008) predict that this "cutting edge" technology will be able to sequence the entire human genome in less than one minute and under a dollar thereby being a serious contender for the Archon X-Prize competition.

DNA based methods for identifying genetic resources

The platforms discussed above may also be used to evaluate specific regions of the genome of a biological entity to determine to which genus, species, or strain it belongs. Herbert posits that:

> microgenomic identification systems, which permit [the discrimination of species] through the analysis of a small segment of the genome, represent one extremely promising approach to the diagnosis of biological diversity; and that in fact, this concept has already gained broad acceptance among those working with the least morphologically tractable groups, such as viruses, bacteria and protists (2003,p 313).

Until the late 1960s, new species of microorganisms were defined through the subjective appraisal and weighting of phenotypic properties, primarily cellular morphology and growth responses on various sugars and other compounds. Although it was known that genes controlled the utilization of these compounds as sources of energy (either by fermentation or oxidative respiration) and could result in an erroneous definition of species, genetic means for species identification did not really exist (Holtz, 1984).

DNA based identification methods such as DNA reassociation analyses were first used in the 1970s for classification of bacteria. It was the seminal work of Woese and his colleagues (Woese and Fox, 1977) that led to the first phylogenetic classification of prokaryotes (a contentious but commonly used name to group *bacteria* and *archaea* together based on their absence of a nucleus; a feature found in eukaryotes) based on the comparison of the nucleotide sequence of the 16S rRNA gene. This gene is universally distributed, highly conserved, evolves very slowly, and plays a key structural role in the ribosome which in turn is part of the cellular machinery involved in protein synthesis (Woese *et al.* 1985; Woese, 1987). All life forms, as we know them, possess ribosomes, so according to the early proposals of Pauling and Zukerkandel (1965), the sequence of this molecule could serve as a molecular chronometer, by which the evolution of different species could be traced.

This work by Woese revealed key distinctions between *Bacteria* and *Archeae*, that members of these two groups shared a common ancestor, but formed two deep and very distinct evolutionary lineages. The third lineage, based on this model of evolution, encompasses the eukaryotes (the plants and animals), which include all living forms that posses a membrane enclosed nucleus and organalles (including the mitochondria and chloroplasts, which we now understand evolved from engulfed, obligatly symbiotic alpha-proteobacterial ancestors in separate events, prior to the evolution of an oxygenic atmosphere on earth. (Gray, 1999; Gray *et al.*, 1999). Like *bacteria* and *archaea*, eukaryotes also contain ribosomes, which in turn contain a 18S rRNA and corresponding 18S rRNA gene, which shares many homologous regions with the 16S rRNA gene. Thus, it is possible to making meaningful comparisons of all species based upon a comparison of the sequence of this gene. Additionally, the sequence of the 16S rRNA gene is approximately 1540 nucleotides, thus there is sufficient information content to provide very far reaching comparisons.

This discovery by Woese has led to a radically different understanding of the evolutionary history of all life and is generally well accepted by contemporary microbiologists, who have abandoned alternative models of classification such as Whittaker's five kingdoms, for which there is little support. In the ensuing 25 years 16S rRNA sequence analysis has since become the principal method by which all *Bacteria* and *Archeae* are classified (Garrity and Holt, 2001; Clarridge, 2004; Garrity and Lilbum, 2005; Cole *et al.*, 2008).. There have been many taxonomic rearrangements, along with nomenclatural changes, to bring earlier classifications into line with the 16S RNA phylogenetic model (a gene tree) and those models are generally well supported by other lines of evidence. Radical changes in sequencing methodologies (*e.g.* invention and commercial development of the polymerase chain reaction, development of automated Sanger sequencing machines) of have greatly accelerated the process.

The 16S rDNA sequence has hypervariable regions, where sequences have diverged over evolutionary time. These are often flanked by strongly-conserved regions. During the sequencing process, PCR primers are designed to bind to conserved regions and amplify variable regions. The DNA sequence of the 16S rRNA gene has been determined for an extremely large number of strains representing the cultivable (28) and non-cultivable phyla (~50). The Ribosomal Database Project (RDP) maintains “archaeal and bacterial small subunit rRNA sequences” (Cole *et al.*, 2008). As of January 14, 2009 the RDP held 759,837 aligned partial 16S rRNA sequences.

It is expected that such changes will continue into the foreseeable future, as genomic methods are brought to bear on the classification of *Bacteria* and *Archeae*. Pilot studies are currently underway to produce complete genome sequences of the taxonomic types of all the major lineages with the ultimate goal of sequencing complete genomes of all of the taxonomic type strains. This is possible because the speed of technical developments in sequencing technology as well as the sharp reduction of cost in producing sequence data.

Analogous developments are currently underway in the fields of botany and zoology. Identifying species of organisms by short sequences of DNA has been at the center of ongoing discussions under the terms DNA barcoding or DNA taxonomy. In an effort to standardize the approach to species identification using molecular techniques it has been proposed that as many species as possible be characterized for the same genetic markers (Blaxter 2004) with the focus being on a 658 base pair fragment of the mitochondrial gene, cytochrome *c* oxidase subunit I (*cox1*) (Hanken 2003; Hebert *et al.,* 2003). This fragment was chosen because it can be found across multiple divergent taxa Thus a sequence of several hundred nucleotides in length acts as a unique identifier for members of a given species, hence the analogy to a computerized barcode label. This method is commonly referred to as the DNA Barcoding approach. There is an international initiative know as The Consortium for the Barcode of Life (CBOL) which is focused on "developing DNA barcoding as a global standard for the identification of biological species" (Frézal and Leblois, 2008).

It is now well understood that a single gene, while highly useful for devising a working taxonomic classification, may not be adequate to yield an accurate identification and that additional gene sequences along with other data may be required to resolve questions of identity at the strain or species level. For example, Vences and colleagues (2005) discovered that the mitochondrial 16S rRNA gene is better suited than the *cox1* gene to serve as a universal DNA barcoding marker in amphibians. Confounding issues include non-uniform distribution of sequence dissimilarity among different taxa and instances in which multiple copies of the 16S rRNA gene may be present in the same organism that differ by more that 5% sequence dissimilarity (Will *et al.*, 2005; Rubinoff *et al.*, 2006; Song *et al.*, 2008). Furthermore, at least in bacteria, it is now known that multiple copies of rRNA genes occur on the same genome and can differ by more than 3.5%, a value that is typically considered to be indicative of membership in different genera (Lagesen, 2008).

Issues

Reducing cost while maintaining—or improving—the quality, length and quantity of the output is a key driver in the design of any new sequencing approach. (Chan, 2005) As mentioned previously, the initial draft of the human genome sequence by Celera (the former commercial arm of the J. Craig Venter Institute) cost an estimated $300 million dollars annually for twenty years. The total cost of the final draft and all the technology that made it possible was estimated to be $3 billion dollars. In 2006 a draft of the genome sequence of the rhesus macaque was completed for $22 million (Service, 2006).

With the ability to sequence much more DNA using the NGS systems, many centers do not yet have the necessary computational hardware and expertise for data analysis and storage. Holt (2008) notes that "any of these new machines running at full capacity for a year will generate, in raw DNA, more sequence than existed in the whole of the National Center for Biotechnology Information (NCBI)–GenBank database at the beginning of 2008". Analysis of the sequence data has rapidly become the bottleneck and developing the necessary skills set and tools will most likely be a driving factor in the execution of next generation sequencing (Lin *et al.,* 2007). Clearly, the currently popular vision of an investigator with a single bench top sequencing machine in place of the current sequencing centers can only be realized with parallel breakthroughs in high-throughput, accessible bioinformatic methods (Hutchinson, 2007).

Recognizing the increasingly urgent need to make genomic data available in standard electronic format scientists and researchers around the world joined forces in 2005 to create the Genomic Standards Consortium (GSC). The primary purpose of this open-membership international working body is to "promote mechanisms that standardize the description of genomes and the exchange and integration of genomic data". Currently this group is working to develop a standard set of core descriptors for genomes and metagenomes known as the Minimum Information about a Genome Sequence (MIGS) and Minimum Information about a Metagenome Sequence (MIMS) specification. (Field *et al.,* 2008).

Future possibilities

It is generally expected that sequencing methods will soon be viewed much in the same way as a microarray experiment is today. In fact, sequencing methodology will likely supplant microarray methods in many situations and whole-genome sequencing will have a host of new applications, such as genotype association studies, mutation screening, evolutionary studies and environmental profiling (Hall, 2007; Kahvejian, 2008). Personalized medicine, based on the genome sequence of an individual and their microbial flora, will become feasible (Pennisi, 2007). To facilitate this movement, in 2003 the NHGRI created the Encyclopedia of DNA Elements (ENCODE), a public consortium focused on identifying all functional elements in the human genome sequence (Weiss, 2007). In 2007, the NIH launched the Human Microbiome Project (HMP), with the goal of identifying the human microbiome as "the collective genomes of all microorganisms present in or on the human body". According to the NIH, the HMP "will lay a foundation for efforts to explore how complex communities of microbes interact with the human body to influence health and disease." This same approach could be easily replicated for other complex interactions between hosts and their associated microbial flora.

Turnbaugh *et al.* (2007) posit that, at a minimum, the following must occur in order for the HMP to be successful:

1. Sequence more reference genomes;
2. Develop an accurate and scalable way to classify the huge numbers of short sequence reads;
3. Develop an initial set of reference microbiomes from healthy individuals;
4. Develop new procedures and increased capabilities for depositing, storing and mining different data types.

In October 2008 the first round of awards for the HMP were awarded for research projects that focused on at least one of the following areas: the development of innovative technologies and computational tools; coordination of data analysis; and the examination of ethical, legal and social implications of such research (Spencer, 2008). Genome-wide association studies in which researchers are comparing the distribution of SNPs in people with and without a particular disease are being conducted to determine how much increased risk is associated with each SNP (Pennisi, 2007). These new technologies are already being used to explore the vast microbial diversity in the natural environment (Tringe *et al.,* 2008) and the untapped genetic variation that can occur in bacterial species. It is expected that these powerful new methods will open up new questions for genomic investigation and will allow high-throughput sequencing to be used for more than just a discovery exercise. Sequence analysis will become a routine assay for hypothesis testing in all disciplines in the life sciences. Some examples are discussed below.

Increasing global concern about the world's bodies of water, improved computer technology and the development of high tech tools have all served to stimulate oceanic research. In the Adriatic Sea, the correlation between mucilage phenomena and the presence of the dinoflagellate *Gonyaulax fragilis* has been recently demonstrated. Tinti *et al.*(2007) developed species-specific molecular probes to monitor microalgal species in seawater samples.

In their work on assessing phytoplankton diversity, Metfies and Medlin (2008) developed a DNA phylochip and transferred 18S rRNA probes from dot-blot or fluorescent in-situ hybridization (FISH) to a microarray format.

The Phylochip is a custom Affymetrix microarray developed by DesSantis *et al.* (2007) at Lawrence Berkeley National Laboratory (LBL) that is currently being used to test water samples in parts of California. This is a microarray that uses genetic probes containing oligonucleotides on the chip to match 16S rRNA gene sequences in waterborne bacteria and archaea. It can detect up to 32, 000 unique versions of the 16S RNA gene found in all bacteria and can be used for identification purposes.

In their efforts to obtain an overview on the occurring microbial community in the soil of a former uranium mine in Germany, Reinicke and Kothe (2008)developed a phylochip, that distinguishes between taxonomic classes.

With the advent of novel micro-fluidics and "lab on a chip" (LOC), it is possible to envision in the next few years the development of a small, inexpensive chip that can be routinely used in the field to identify unknown bacterial species (Mauk *et al.*, 2007; Bjerketorp *et al.*, 2008. Recently technology for making chips from paper and double-stick tape have shown the possibility of making these for a few pennies each. (Grant, 2008; Martinez *et al.*, 2008).

Discussion

It is clear to us that very low cost sequencing technology along with sophisticated bioinformatics tools will soon be available to presumptively identify a genetic resource, with some degree of accuracy and reliability, at the point of need. Development of the underlying technology is being driven by developments outside the field of biodiversity research, however; those tools and technologies will be readily adaptable to meet the anticipated needs of the IR. We strongly advocate the adoption of technologies that are open, robust, widely accepted and proven (*e.g.*, standards) rather than creating specialized alternatives. Such a strategy will encourage widespread adoption and lessen the chance of selecting a dead-end approach.

The concept of identification is central to the goals of the CBD ABS regime, which rests on the fundamental principle that a user is legally obliged to share in the benefits obtained from the use of a particular genetic resource with the provider. Under some circumstances, identification to the family, genus or species level may be adequate and identification methods based on a single gene may be the appropriate tool (e.g., biotic inventories, wild-life management, ecological studies). However, there is ample evidence derived from more then a half century of natural product screening that is now supported by extensive genomic data that such approaches are grossly inadequate in cases where the traits of interest may be found only in subpopulations or individuals within a species or the same trait is found to be widely distributed across taxa as a result of horizontal gene exchange. Any tracking system that is implemented must accurately reflect current knowledge and readily incorporate new knowledge, presumably over a long time frame as transactions involving genetic resources may be long lived (>20 yrs).

The number of items to be identified (or "tag") is a challenge and the extent of the task will depend largely on the legally required "granularity" of the identification. One could approach this as a taxonomic and nomenclatural problem, however, this would result in an open-ended task as the finer the granularity required the greater the number of biological items to be identified, accompanied by constantly shifting taxonomic boundaries and accompanying nomenclatural changes and an ever expanding mass of data. Without careful attention and forethought to the management and storage of this data, trying to retrieve information about a particular biological item could become like trying to find a needle in a haystack. An alternative approach is to define lightweight metadata models that adequately define the type of genetic resources and associated information that must be bound to those genetic resources to facilitate CBD ABS objectives, and make the resulting information objects readily accessible using a persistent identifier system.

The next section includes a discussion of how persistent identifiers are used; briefly reviews six of the most commonly used identifiers; addresses the issues that need to be examined when selecting a persistent identifier scheme, and offers suggestion for implementation of a persistent identifier scheme that will facilitate ABS.

References

Archon (2008) Prize Overview: A $10 million prize for the first team to successfully sequence 100 human genomes in 10 days. Available at: http://genomics.xprize.org/genomics/archon-x-prize-for-genomics/prize-overview

Bjerketorp, J., Ng Tze Chiang,A., Hjort,K., Rosenquist, M., Liu,W., and Jansson, J. (2008) Rapid lab-on-a-chip profiling of human gut bacteria. *Journal of Microbiological Methods,* 72(1): 82-90. .doi:10.1016/j.mimet.2007.10.011

Blaxter, M. L. (2004.The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences,* 359(1444): 669-679. doi: 10.1098/rstb.2003.1447

Blow, N., 2008.DNA sequencing: generation next-next. *Nat Meth,* 5(3): 267-274. doi: 10.1038/nmeth0308-267

Branton, D., .Deamer, D., Marziali, A., Bayley, H., Benner, S., Butler, T. *et al*. (2008). The potential and challenges of nanopore sequencing. *Nat Biotech,* 26, (10): 1146-1153. doi: 10.1038/nbt.1495

Brent, M., 2008. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat Rev Genet,* 9 (1): 62-73. doi: 10.1038/nrg2220

Broeder, D. (2007) *Persistent Identifiers*" presented at the DAM-LR Meeting, Lund University, December 18, 2007. http://www.mpi.nl/DAM-LR/meeting5/Persistent Identifiers.pdf

Chaisson, M., Pevzner, P. and Tang, H. (2004) Fragment assembly with short reads. *Bioinformatics,* 20(13): 2067-2074. doi:10.1093/bioinformatics/bth205

Chan, E. (2005) Advances in sequencing technology. *Mutation Research,* 573 (1-2): 13-40. doi: 10.1016/j.mrfmmm.2005.01.004.

Clarridge, J. (2004) Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clin. Microbiol. Rev.,* 17(4): 840-862. doi: 10.1128/CMR.17.4.840-862.2004.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. *et al*. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucl. Acids Res,* 37(Database issue), 141-145. doi:10.1093/nar/gkn879

Desantis, T., Brodie, E., Moberg,J., Zubieta, I., Piceno, Y. and Andersen, G. (2007) High-Density Universal 16S rRNA Microarray Analysis Reveals Broader Diversity than Typical Clone Library When Sampling the Environment. *Microb Ecol*, 53 (3): 371-383. doi: 10.1007/s00248-006-9134-9

Faris, S. M., and .Eakin, J. (2008) *What will it take to sequence the human genome error-free, in minutes, for pennies?* Reveo. Available at: http://www.lab-robotics.org/Upstate/Archives/071105 material/LRIG Presentation-Reveo.pdf

Field, D., Garrity, G., Gray,T., Morrison, N., Selengut,J., Sterk, P., *et al*. (2008)The minimum information about a genome sequence (MIGS) specification. *Nature biotechnology,* 26(5). 541–547. doi: 10.1038/nbt1360

Fleischmann, R., Adams, M., White, O., Smith, H., and Venter, J. (2003) Nucleotide sequence of the *Haemophilus influenzae* Rd genome, fragments thereof, and uses thereof, US Patent 6528289.

Fleischmann, R., Adams, M., White, O., Clayton, R. Kirkness, E., Kerlavage, A. *et a*l. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269 (5223): 496-512. 10.1126/science.7542800

Fraser, C., Eisen, J. Nelson, K., Paulsen, I., and Salzberg, S. (2002) The Value of Complete Microbial Genome Sequencing (You Get What You Pay For). *J. Bacteriol.,* 184 (23): 6403-6405. doi: 10.1128/JB.184.23.6403-6405.2002

Fredlake, C., Hert, D., Kan, D., Chiesl, T., Root, B., Forster, R. *et al*. (2008) Ultrafast DNA sequencing on a microchip by a hybrid separation mechanism that gives 600 bases in 6.5 minutes. *Proceedings of the National Academy of Sciences of the United States of America* 105 (2): 476–481. doi: 10.1073/pnas.0705093105

Garrity, G. M., and Holt, J. (2001) The road map to the Manual. In G. Garrity, D. Boone, and R. Castenholz (Eds.), *Bergey's Manual of Systematic Bacteriology* (2nd ed.,Vol 1 pp 119-166). Springer.

Garrity, G.M., Field, D., Kyrpides, N., Hirschman, L., Sansone, S., Angiuoli, S., *et al.,* (2008) Toward a Standards-Compliant Genomic and Metagenomic Publication Record. *OMICS: A Journal of Integrative Biology,* 12(2): 157-160. doi: 10.1089/omi.2008.A2B2

Garrity, G. M. (2007) *An Overview of Persistent Identifiers*. presented at the IT Support for SMTA implementation , Rome Italy, February 14, 2007.

Garrity, G. M., and Lilburn, T. (2005) Self-organizing and self-correcting classifications of biological data. *Bioinformatics,* 21(10): 2309-2314. doi:10.1093/bioinformatics/bti346

Grant, B. (2008) The 3 cent microfluidics chip. *The Scientist* (December 8, 2008). http://www.the-scientist.com/news/display/55270

Gray, M. W. (1999). Evolution of organellar genomes. *Curr Opin Genet Dev* 9, 678-687. doi:10.1016/S0959-437X(99)00030-1

Gray, M.W.,Burger, G and Lang, B.F. (1999) Mitochondrial Evolution, *Science* 283: 1476-1481. doi: 10.1126/science.283.5407.1476

Gupta, P. (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology* 26(11) (November 2008): 602-11. doi:10.1016/j.tibtech.2008.07.003

Hall, N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology,* 210(9): 1518-1525. doi: 10.1242/jeb.001370

Hanken, J. (2003) *DNA Barcoding.* Presentation at the NY Botanical Garden, New York on Dec 11, 2003. http://www.barcoding.si.edu/presentations/Hanken_barcoding_nybg_11dec.pdf

Hayden, E. (2008) International genome project launched. *Nature,* 451: 378-379. doi:10.1038/451378b

Hebert, P., Cywinska, A., Ball, S., and deWaard, J. (2003) Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B,* 270: 313-321. doi: 10.1098/rspb.2002.2218

Hillier, L., Gabor, G. Quinlan, A., Dooling, D., Fewell,G., Barnett, D., *et al.* (2008) Whole-genome sequencing and variant discovery in C. elegans. *Nature Methods*, 5(2): 183-188. doi:10.1038/nmeth.1179

Hilse, H., and Kothe, J. (2006) *Implementing Persistent Identifiers: overview of concepts, guidelines and recommendations*. European Commission on Preservation and Access (ECPA) report 18. Available at: http://www.knaw.nl/ecpa/publ/pdf/2732.pdf

Holt, R., and Jones, S. (2008) The new paradigm of flow cell sequencing. *Genome Research,* 18 (6): 839-46. doi: 10.1101/gr.073262.107.

Hutchison, C. (2007) DNA sequencing: bench to bedside and beyond. *Nucl. Acids Res.*, 35 (18): 6227-6237. doi:10.1093/nar/gkm688

InParanoid (2008) http://inparanoid.sbc.su.se/cgi-bin/summary.cgi

Kahn, R., and Wilensky, R. (1995) *A Framework for Distributed Digital Object Services*: Technical Report tn95-01. Corporation for National Research Initiatives. Available at: http://www.cnri.reston.va.us/k-w.html. (re-published, with an additional introduction by the authors in International Journal on Digital Libraries (2006) 6(2): 115-123. doi: 10.1007/s00799-005-0128-x

Kahvejian, A., Quackenbush, J., and Thompson, J.(2008) What would you do if you could sequence everything? *Nat Biotech,* 26 (10): 1125-1133. doi: 10.1038/nbt1494

Käller, M., Lundeberg, J. and Ahmadian, A.(2007) Arrayed identification of DNA signatures. *Expert Review of Molecular Diagnostics,* 7 (1): 65-76. doi:10.1586/14737159.7.1.65

Karow, J. (2008) PacBio to Start Selling Next-Gen Sequencer to Early Users in 2010; Goal is 100 Gb/Hour. *In Sequence*, 2 (7). Available at: http://www.in-sequence.com/issues/2_7/webreprints/145257-1.html

Kasianowicz, J., Brandin, E., Branton, D., and Deamer, D. (1996) Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America,* 93(24): 13770-13773. doi: 10.1073/pnas.93.24.13770

Kunze, J. (1996) *Persistent Identifier Principles and Practice* presented at the International Conference on Dublin Core and Metadata Applications, Berlin, September 24, 2008. Available at: http://dc2008.de/wp-content/uploads/2008/09/dc2008_id_all_slides.pdf

Lagesen, K. (2008) *Computational characterization of ribosomal RNAs in prokaryotes*. Doctoral dissertation. University of Oslo, Oslo, Denmark.

Levene, M., Korlach, J., Turner, S., Foquet, M., Craighead, H., and Webb, W. (2003) Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science,* 299 (5607): 682-686. doi: 10.1126/science.1079700

Levy, J. (1994) Sequencing the yeast genome: An international achievement. *Yeast*, 10 (13): 1689-1706. doi: 10.1002/yea.320101304

Lin, F., Schröder, H., and Schmidt, B. (2007) Solving the Bottleneck Problem in Bioinformatics Computing: An Architectural Perspective. *The Journal of VLSI Signal Processing,* 48 (3): 185-188. doi: 10.1007/s11265-007-0088-z

Liolios, K., Mavromatis, M., Tavernarakis, N., and Kyrpides, N. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucl. Acids Res.,* 36 (Database issue): D475-479. doi:10.1093/nar/gkm884

Lorenz, J., Jackson, W., Beck, J., and Hanner, R. (2005) The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 360 (1462): 1869-1878. doi: 10.1098/rstb.2005.1718

Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437 (7057): 376-80. doi:10.1038/nature03959

Marieke , G. (2005) *An Introduction To Persistent Identifiers QA Focus* briefing document no. 80. UKOLN. Available at: http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-80/html

Martinez, A., Phillips, S., and Whitesides. G. (2008) Three-dimensional microfluidic devices fabricated in layered paper and tape. *Proceedings of the National Academy of Sciences*, 105(50): 19606-19611. doi: 10.1073/pnas.0810903105

Mauk, M., Ziober, B., Chen, Z., Thompson, J., and Bau, H. Lab-on-a-Chip Technologies for Oral-Based Cancer Screening and Diagnostics: Capabilities, Issues, and Prospects. *New York Academy Sciences Annals,* 1098: 467-475. doi: 10.1196/annals.1384.025

Metfies, K., and Medlin, L. Feasibility of transferring fluorescent in situ hybridization probes to an 18S rRNA gene phylochip and mapping of signal intensities. *Applied and environmental microbiology*,74 (9): 2814-21. doi:10.1128/AEM.02122-07

Paskin, N. (1999) Toward unique identifiers. *Proceedings of the IEEE*, 87 (7): 1208-1227. doi: 10.1109/5.771073

Pennisi, E. (2007) Breakthrough of the Year: Human Genetic Variation. *Science,* 318 (5858): 1842-1843. doi: 10.1126/science.318.5858.1842

Pennisi, E. (1997) "Genome Sequencing: Microbial Genomes Come Tumbling In." *Science,* 277(5331): 1433. doi: 10.1126/science.277.5331.1433

Qiu, J., and Hayden, E. (2008) Genomics sizes up. *Nature,* 451(234) doi:10.1038/451234a

Quail, M.A. *et al.*, 2008. A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, 5(12), 1005-10. doi:10.1038/nmeth.1270

Reinicke, M., and Kothe, E. (2008) Microbial monitoring and population dynamics in AMD affected field site. *Geophysical Research Abstracts* 10, EGU2008-A-06985. Available at : http://www.cosis.net/abstracts/EGU2008/06985/EGU2008-A-06985.pdf

Reveo (2008) Reveo, Inc. registers to compete in the $10M Archon Genomics X-Prize. Retrieved from http://www.reveo.com/us/node/46

Ronaghi, M. (2001) Pyrosequencing Sheds Light on DNA Sequencing. *Genome Research,* 11: 3-11. doi: 10.1101/gr.150601

Rubinoff, D., Cameron, S., and Will, K.A Genomic Perspective on the Shortcomings of Mitochondrial DNA for "Barcoding" Identification." *Journal of Heredity,* 97(6): 581-594. doi: doi:10.1093/jhered/esl036

Salamone, S. (2002) LSID: An Informatics Lifesaver. *Bio-itworld.com*, January 12, 2002. http://www.bio-itworld.com/archive/011204/lsid.html

Sanger, F., Nicklen, S., and Coulson.A. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Nati. Acad. Sci. USA* 74(12): 5463-5467

Schuster, S. (2008) Next-generation sequencing transforms today's biology. *Nature Methods,* 5(1): 16-18. doi:10.1038/nmeth1156

Service, R. (2006) Gene Sequencing: The Race for the $1000 Genome. *Science,* 311(5767): 1544-1546. doi: 10.1126/science.311.5767.1544

Shendure,J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotech,* 26(10): 1135-1145. doi: 10.1038/nbt1486

Siva, N. (2008) 1000 Genomes project. *Nat Biotech,* 26(3): 256. doi: 10.1038/nbt0308-256b

Smith, D., Quinlan, A., Peckham, H., Makowsky, K., Tao, W., Woolf, B., *et al.* (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research,* 18(10): 1638-1642. doi: 10.1101/gr.077776.108.

Song, H., Buhay, J., Whiting, M., and Crandall, K. (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America,* 105 (36): 13486-91. doi: 10.1073/pnas.0803076105

Storm, A., Chen, C. Zandbergen, H. and Dekker, C. (2005) Translocation of double-strand DNA through a silicon oxide nanopore. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics,* 71(5 Part 1): 051903. doi: 10.1103/PhysRevE.71.051903

Storm, A., Storm, C., Chen, J., Zandbergen, H., Joanny, J.,and Dekker,C. (2005) Fast DNA translocation through a solid-state nanopore. *Nano Letters*,5 (7): 1193-11977. doi: 10.1021/nl048030d

Ten Bosch, J. and Grody, W. (2008) Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *The Journal of Molecular Diagnostics,* 10 (6): 484-92. doi: 10.2353/jmoldx.2008.080027

Tinti, F., Boni,L., Pistocchi, R., Riccardi,M., and Guerrini, F. (2007) Species-specific probe, based on 18S rDNA sequence, could be used for identification of the mucilage producer microalga *Gonyaulax fragilis* (Dinophyta). *Hydrobiologia,* 580 (1): 259-263. doi: 10.1007/s10750-006-0446-z

Tonkin, E. (2008) Persistent Identifiers: Considering the Options. *Ariadne*, no. 56 (July 10, 2008). Available at: http://www.ariadne.ac.uk/issue56/tonkin/

Tringe, S. and Hugenholtz, P. (2008) A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology,* 11(5) (October 2008): 442-446. doi:10.1016/j.mib.2008.09.011

Turnbaugh, P., Ley, R., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon. J. (2007) The Human Microbiome Project. *Nature,* 449 (7164): 804-810. doi: 10.1038/nature06244

Ussery, D., and Hallin, P. (2004) Genome Update: annotation quality in sequenced microbial genomes. *Microbiology,* 150, 2015-2017. doi: 10.1099/mic.0.27338-0

Vences, M., Thomas,M., van der Meijden,A., Chiari,Y. and Vieites, D. (2005) Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Frontiers in Zoology,* 2(1): 5. doi: 10.1186/1742-9994-2-5

Venter, J.C., Adams, M., Myers, E., Li, P., Mural, R., Sutton, G., *et al*. (2001) The Sequence of the Human Genome. *Science,* 291(5507) (February 16, 2001): 1304-1351. doi: 10.1126/science.1058040

Visigenbio (2008) Press Releases VisiGen Receives Patent for Real-Time Single-Molecule DNA Sequencing http://visigenbio.com/press_patent_feb_08.html

Von Bubnoff, A. (2008) Cell : Next-Generation Sequencing: The Race Is On. *Cell,* 132 (5), (March 2008): 721-723. doi:10.1016/j.cell.2008.02.028

Weiss, R. (2007) Intricate Toiling Found In Nooks of DNA Once Believed to Stand Idle. *Washington Post,* June 14, 2007, F edition.

Will, K., Mishler, B. and Wheeler, Q. (2005)The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology,* 54(5) (October 2005): 844-51. doi: 10.1080/10635150500354878

Woese, C R, and Fox, G.(1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America,* 74(11) (November 1977): 5088-90

Woese, C R, Stackebrandt, E., Macke,T. and Fox, G. (1985) A phylogenetic definition of the major eubacterial taxa. *Systematic and Applied Microbiology,* 6, 143-51.

Woese, C R, L J Magrum, and Fox, G. (1978) Archaebacteria. *Journal of Molecular Evolution,* 11(3): 245-51. doi: 10.1007/BF01734485

ZSG (2008) ZSG process enables electron microscopes to "see" DNA. Retrieved from: http://zsgenetics.com/application/GenSeq/commercial.html

Zuckerkandl, E and Pauling, L. (1965) Evolutionary Divergence and Convergence in Proteins. In V. Bryson and H.Vogel (Eds.), *Evolving Genes and Proteins* (pp. 97–166) New York: Academic Press.

PERSISTENT IDENTIFIERS

The rapid progress in digital technology has led to an explosion of global information storage on, and dissemination via, the Internet (notably the World Wide Web (WWW), though this is not the only internet application likely to be of interest for future internet information management). For example, ten major natural history museum libraries, botanical libraries, and research institutions are currently digitizing over two million volumes of biodiversity literature held in their respective collections in order to make this information available for open access at the Biodiversity Heritage Library and to serve as part of a global "biodiversity commons." The resulting digital archive will be available to anyone anywhere in the world who has access to a computer, the internet and the world wide web. The issue arises however in that information needs to be displayed AND to be uniquely identified or referenced in such a way that it can be consistently and readily retrieved over time in a dynamic digital information environment. One way in which to address both needs is through the use of unique identifiers that are actionable (*i.e.*, resolvable or "de-referenceable").

Before the onset of the global digital era, one of the best known identifiers in commercial use for content was the International Standard Book Number (ISBN), a unique, numerical commercial book identifier (Marieke, 2005). Each edition and variation (except reprinting) of a book is given a unique ISBN. The ISBN is part of an extending family of separate International Organization for Standardization (ISO) identifiers covering a range of "information and documentation" types (ISO TC46/SC9); with digital convergence the need for these separate identifiers to become interoperable has begun to be recognized (Paskin 2006).

On the World Wide Web (the Web), documents have been, and are currently, accessed through the use of Uniform Resource Locators (URLs), which are used to create hyperlinks on the web. Because the access method – viewing or requesting documents via the Internet – required the use of a URL as addressing mechanism, the URL has also become a common way of referencing documents, *i.e.*, in a citation (Hilse *et al.*, 2006)[3]. A URL provides fast, direct access to the document as long as it does not change location (*e.g.*, moved to a different server or a different location on the same server) and the contents of the document remain unchanged. Tonkin (2008) explains it thusly:

> URLs are often implemented using the server's file system as a kind of lookup database: for example, *http://www.ukoln.ac.uk/index.html* is a file called 'index.html' that is situated in the root directory of the Web server running on port 80 of the machine responding to *www.ukoln.ac.uk*. Because it was very simple to get up and running quickly, many early servers tended to refer to digital objects in this way. For example, *while http://www.ukoln.ac.uk/index.html* means "The digital object that is provided when you ask the Web server (running on port 80 of the IP [internet protocol] that is currently returned by a Domain Name System (DNS) server for the domain *'www.ukoln.ac.uk'*) about the string

[3] Owing to the problems of URL citation discussed below, formal publication citations are increasingly making use of other forms of (persistent) identifier as described later in this paper.

> '/index.html'", for many servers this has meant "The file named ‘index.html’ at the root level". There are certain advantages to this approach: for example, a clear mapping between the file system and the structure of the Web site can make it easier for maintainers to understand how the site is structured, but with no other mechanism in place, if someone removes the file the link will break – easy come, easy go.

The reality is that an object (which could be a document, a photo, a software program, data, *etc.*) may be moved, removed, duplicated elsewhere, revised or renamed for any number of reasons. This lack of persistence, commonly referred to as “linkrot” leads to 404 errors (file not found). This would be akin to a book in a very large library not being on the shelf indicated in the catalogue. When confronted with such a situation the only solution for the end user is to use additional metadata (such as the author or title in the case of a document) to look for the object, using search engines such as Google or specialized information, *e.g.* about the institution (university, library, publisher, *etc.*) that owns the object. Linkrot inhibits access to objects and causes problems when archiving material for long-term preservation and permanent access. Along the same lines, there are multiple copies of "the same book" residing in numerous libraries; each in its own particular location. Each book has associated with it its own metadata, particular to that institution.

The transience of websites, as well as the documents found on the web, causes serious problems for end users, especially for organizations and institutions that are working to integrate these documents into their local systems. It is not at all unusual to find URLs in databases such as online catalogs that are no longer valid. The source of this problem is that the URL was created to designate the *location* of objects on the web. It was never meant to serve as a digital identifier for objects on the web.

To overcome the problems arising from the use of URLs as both a locator and identifier of objects, focus has been placed on designing and implementing a system based on an identifier that would be assigned to an object when it was created, or even before that, and that would remain persistently associated with that object (Kahn and Wilensky, 1995). In other words, the object would be assigned an identifier that would be unique and persistent and independent of the location of the object. According to the International Digital Object Identifier (DOI) Foundation (http://www.doi.org) there are two principles to such persistent identification

1. Assignment of an identifier to an object: Once assigned the identifier must identify the same object beyond the lifetime of the object or identifier
2. Assignment of an object to an ID: explicitly stating and defining “what is” the referent of the identifier.

Paskin (2004) contends that the most useful persistent identifiers are also persistently actionable (that is, you can "click" them, like a URL, to link the identifier to some service); however, unlike a simple hyperlink, persistent identifiers are supposed to continue to provide access to an object, even when it moves to other servers or even to other organizations. Paradigm (2008) suggests that the locations of objects may change even more frequently owing to the need for regular refreshment of storage media to guard against media failure. It is also likely that an instance of intellectual property entity acquired in digital form will become associated with multiple representations of itself over time, as new technologies result in different storage formats than those used to store the original object (Paradigm, 2008). The persistent identifier improves discovery and access over time, despite place and technology dislocation.

An important strategy to help reduce the danger of being unable to retrieve an object is to add a managed layer of indirection between the browser and the target object. The persistent identifier itself references an object which may be available in multiple locations or manifestations, rather than referring to a specific instance or representation of the object. Retrieving an instance requires a resolver that forwards the end user to a current (or local) instance of the object. A resolver database maps the location of the object and redirects the user to a current location. The resolver service is intended to redirect the browser to an appropriate or current copy of the object. Indirection is typically transparent to the end user as shown in **Figure 3**.

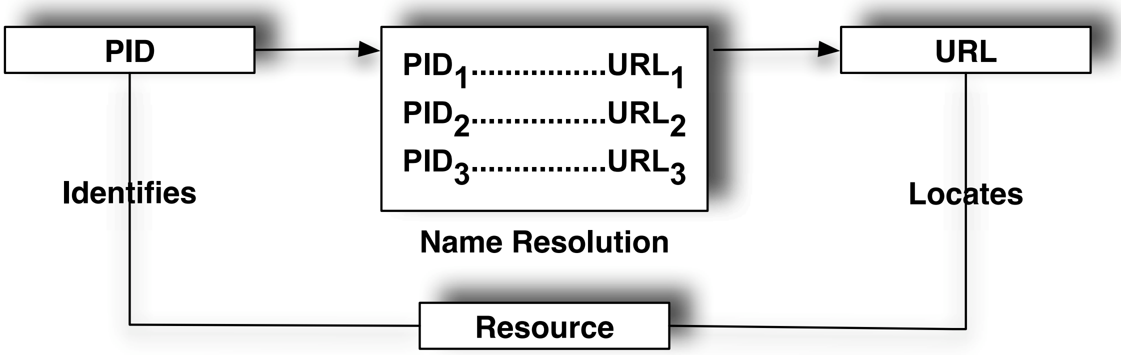

Figure 3 General model of name-to-location resolution using persistent identifiers. Digital content and other identifiable objects are permanently assigned a persistent unique identifier string (preferably one that contains no mnemonic information). The persistent identifier identifies an object on the Internet such as an article or part of an article, an image or some other useful piece of information that is likely to be retrieved for various reasons. Each object is found on one or more servers at given location that is subject to change over time. The mapping of the PID in a query to the location of the object on the Internet is handled by a resolver. When the PID is used to identify an object rather than a URL, the former value can be used with a high degree of reliability, without any further modification, so long as the mapping of URLs is regularly updated and checked. Various implementations of name resolution are discussed below and are based on this underlying conceptual model.

Persistent identifiers (PIDs) are widely used to provide access to publications and data and are unaffected by the changes in storage and services that data might undergo over its lifespan. PIDs also allow access to the metadata (*i.e.*, the descriptive information about the structure and contextual meaning of the data). Creating persistent identifiers, and keeping the associated metadata and storage information current and useable is a critical part of the responsibility for the long-term curation of research data. Clearly, persistent identifiers should be an integral part of an organization's comprehensive data management strategy.

Persistent Identifier Schemes

In the digital environment, PIDs must be unique, persistent, and first class[4] (*e.g.*, independent of potentially changeable specific Web domain names). In addition, they must be resolvable using currently deployed common technology (*e.g.*, standard Web protocols, such at http), and flexible enough to allow efficient management of digital information and accommodate technology improvements, and applicable to a wide universe of objects. They must also be based on social and technical infrastructures that ensure longevity far into the future. Over the last ten years, a number of persistent identifier schemes have been developed by different communities to meet their specific needs. An institution or sector may find that some elements from a number of different schemes are suitable for its use rather than any one scheme in its entirety; or it may find that none of the existing schemes are appropriate, resulting in the need to develop its own solution (in which case, the disadvantages of potential lack of interoperability with other schemes must be considered).

A number of identifiers are currently in use within the life sciences such as the INSDC identifiers (e.g., sequence accession numbers used at GenBank, EMBL, and DNA Database of Japan) and project IDs associated with specific genome sequencing efforts as well as a number of unique institutional identifiers associated with samples and specimens. However, these are largely institution specific, *i.e.*, used only within the institutions for which they were created, or are controlled by those organizations. A PubMed ID, which is used to identify content within the National Library of Medicine, is an example of such an institutional identifier.

Three terms that are used frequently in discussions about identifiers and about which there may be some confusion, are Uniform Resource Identifier (URI), Uniform Resource Name (URN) and Uniform Resource Locator (URL). A URI is a string of characters that identifies an object by location, or a name or both. Some examples of URI include:

http://cbd.int
ftp://ftp.is.co.za/
mailto:ABS@cbd.int
telnet://192.0.2.16:80
urn:isbn:3540240225

URLs and URNs are subsets of URIs. A URL is a specialization of URI that defines the network location of a specific object, so in the above example one knows to use the URL *http://cbd.int* to find information about the CBD.

A URN (which will be discussed in greater detail below) is a string of characters that identifies an object by name in a given namespace but it does not include the location of the object.

Unlike a URN, the URL defines how the object can be obtained. A URN is analogous to the name of the person, while a URL is like the street address of that person. The URN defines the identity of a thing while the URL provides a method for finding that thing: essentially, "what" vs. "where".

[4] First class = independent of any other item.

URNs are often compared to the ISBN system for uniquely identifying books (an ISBN can be encoded as a URN as is shown in the example above). Having the unique identifier of a book lets one discuss the book, such as whether one has read it or bought or sold it, *etc*. To actually pick up and read the book, however, one must know where the book is (*e.g.* "it is on the bedside table"). URNs and URLs are often complementary.

The following is an overview and brief discussion of six persistent identifier schemes currently in use across different domains and by a number of different organizations:

Uniform Resource Name (URN)
Persistent Uniform Resource Locator (PURL)
Archival Resource Key (ARK)
Life Science Identifiers (LSID)
Handle System (Handle)
Digital Object Identifier System (DOI)

Uniform Resource Name (URN)

Uniform Resource Name (URN) was created in 1992 to serve as a persistent location independent object identifier (Deutsche Nationalbibliothek, 2001). URNs *do not* describe the location or the availability of the identified object. Standards for the URN are controlled, developed and published in the form of “Request for Comments” (RFCs) by the Internet Engineering Task Force (IETF). URNs are designed to make it easy to map other namespaces (that share the properties of URNs) into URN-space. Although designed to be independent of any underlying technologies such as Domain Name System (DNS), the only present technique of resolving URNs on the Internet is based on DNS. There are no widely standardized ways of use; *i.e.*, you cannot type URNs into a browser except in certain special circumstances (Paskin, 2006).

All URNs have the following syntax (Moats, 1997):

URN:[Namespace Identifier]:[Namespace Specific String]

URN:ISBN:0-395-36341-1

The leading “URN:” sequence is case-insensitive. The Namespace Identifier (NID), which is also case-insensitive, identifies which namespace is being used, and comes from the URN Registry maintained by the Internet Assigned Numbers Authority (IANA). This registry lists existing naming schemes, some of which were created for other than the digital environment, *e.g.* ISBN, International Standard Serial Number (ISSN) and National Bibliographic Numbers (NBN, a namespace assigned to national libraries for integrating different identification schemes into the same identifier namespace). The NID can consist of letters, numbers and hyphens.

The Namespace Specific String (NSS) follows the NID and is preceded by a colon. The NSS can consist of any character and is dependent on the namespace from which it comes (*e.g.* a string of numbers in the case of ISBN); those characters which are outside the URN character set must be encoded using UTF-8 encoding.

URNs were developed to be independent of any one resolution service. A number of different approaches have been proposed, although a universal resolution service for URNs is not yet available. There is also some disagreement within the Web and Internet community as to whether URNs are necessary at all (Paradigm, 2008). These pose an obstacle to their widespread adoption.

To use URNs as persistent identifiers, an organization can either work within an existing URN initiative which has been assigned a NID, or (where a new, globally unique approach to identifiers has been developed) obtain a new NID, through a standardized application procedure. As of December 9, 2008, although there were thirty nine registered formal URN Namespaces listed on the Official IANA Registry of URN Namespaces there were actually thirty five different registrants since four of them have each registered for more than one namespace.

Notable Design Features of URNs and comments

- Since the NSS can take any form, other namespaces can be easily mapped into URNs. This makes URNs flexible and easy to construct. Global uniqueness is preserved as long as the NSS is unique within the NID (Paradigm, 2008).
- The URN is an open standard and technology independent.
- URNs can be used with the DNS and HTTP, meaning that a URN can be coded into a URL, and a proxy server can be used to route URN requests to a host server. This allows URNs to be resolved using a standard web browser (Paradigm, 2008).
- There is no universal resolution service.
- If the NSS is appended to a URL to form a query for a browser, it is no longer a persistent name since the name and location are the same.
- The lack of consensus about the value of URNs puts the long-term future of URNs in doubt. There is also a lack of underlying policy and social/business infrastructure necessary to makes this operational in a sustainable way.

Persistent Uniform Resource Locator (PURL)

The Persistent Uniform Locator (PURL) system was developed by the Online Computer Library Center (OCLC) in the USA to be a library cataloguing application. PURLs were first implemented in 1996 in the Internet Cataloguing Project - which aimed to advance practice and standards for cataloguing internet objects, and addressed the issue of including URLs in cataloguing records. PURL takes the concept of the URL and adds a resolution service layer. Instead of the URL pointing directly to the object in question, it points to an intermediate PURL resolution service. The PURL resolution service is used to look up the actual URL of the object and return that URL to the user. The user can then complete the URL transaction in the normal fashion. This is commonly referred to as HTTP “redirection”, redirecting the user query to the appropriate object and is depicted in **Figure 4**. PURLs are compatible with other document identification standards such as the URN. In this sense, PURLs are sometimes described as an interim solution prior to the widespread use of URNs.

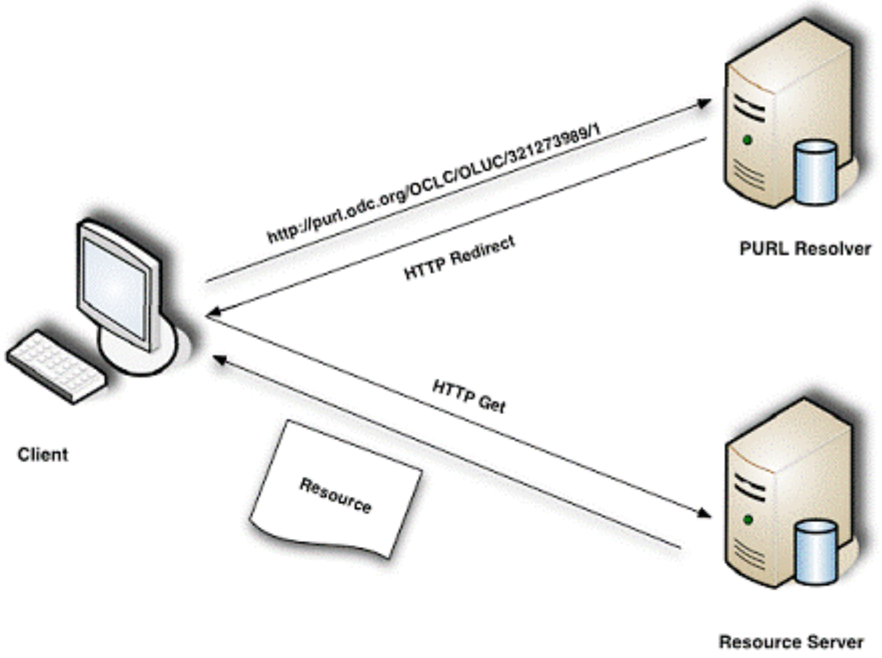

Figure 4. PURL Resolution. PURL identifiers are passed by a client as an argument to the PURL resolver in the form of a URL. The PURL resolver responds by sending the client an HTTP redirect that is then used to form an HTTP GET to retrieve the object of interest from the web server hosting that content.

A PURL contains the URL for the PURL Resolver Service (in the example above, the resolver at OCLC is used) followed by an identifier for the object. Of course, the resolver server must be updated when the actual URL location changes, but the PURL of the document does not change. A PURL server can resolve only the PURLs it maintains (*e.g.*, the PURL server of OCLC cannot resolve PURLs assigned by other PURL servers). The PURL Resolver software is available free from OCLC, or PURLs can be deposited on the OCLC Resolver under an agreement with OCLC.

The syntax of a PURL consists of a protocol, resolver address and the name assigned to an object. The example below was borrowed from www.purl.oclc.org.

http://purl.oclc.org/OCLC/PURL/FAQ

Protocol:	http
Resolver address:	purl.oclc.org
Name:	OCLC/PURL/FAQ

The OCLC PURL server is still up and running at http://purl.oclc.org/ and may be used by anyone. Everyone is invited to establish their own sub-domain on this server and maintain a PURLs. (Hilse and Kothe, 2006). As of January 17, 2009 there had been 728,399 PURLS created and 546,060,979 PURLS resolved (OCLC, 2009)

Notable Design Features of PURLS and comments

- Because it uses existing services PURL is easy and inexpensive to create and resolve (Paradigm, 2008).
- The system is compatible with URI and URN schemes and is standards based.
- PURLs, originally developed to address library cataloguing problems, can provide an effective means of linking from an Encoded Archival Document (EAD) catalogue entry to the associated dissemination information package (DIP).
- PURL is scalable and through the use of the distributed technology of DNS/HTTP, many different PURL servers can be established locally. This means that servers are not overloaded and that there is greater local control over PURL creation (Paradigm, 2008) thereby avoiding the overloading of servers and enabling greater local control over PURL creation.
- PURLs were designed primarily as identifiers for open, web-based objects (essentially ‘published’ material), while digital archives have unique requirements. Repositories for personal digital archives must identify closed or restricted access material and various metadata. Therefore digital repositories need to implement PURLs locally in a way to prevent unauthorized access.
- In a personal digital archive each individual object must be unambiguously identifiable, so a facility like partial resolution is inappropriate.
- PURLs do not require the use of a global directory to ensure globally unique identifiers are used and do not have the underlying infrastructure to ensure persistence.
- There is a lack of underlying policy and social/business infrastructure necessary to makes this operational in a sustainable way.

Archival Resource Key (ARK)

The Archival Resource Key (ARK) was developed by Kunze and Rodgers (2001) at the conclusion of a study of persistent identifier systems for the US National Library of Medicine (NLM) and is maintained at the California Digital Library (CDL) within the University of California. ARK introduces a concept combining the features that a persistent identifier should have and building a technical and administrative framework on that concept. Its focus is on resolving and delivering metadata. An ARK is a URL created to allow persistent, long-term access to information objects. ARKs can identify objects of any type: digital documents, databases, images, software, and websites, as well as physical objects (books, bones, statues, *etc.*) and even intangible objects (chemicals, diseases, vocabulary terms, performances) (Kunze, 2003, 2008; Kunze and Rodgers, 2008).

ARKs supports persistent identification, which is necessary and useful because both the protocols used to access objects (such as http and ftp) and the sites that host the objects are subject to change (Kunze, 2008). An ARK contains parts that are impervious to such changes and parts that are flexible enough to support technological changes.

An ARK connects to three things:

1. the object itself,
2. a brief metadata record when a single question mark is appended to the ARK,
3. a maintenance commitment from the current server when 2 questions mark are appended to the ARK.

The ARK syntax can be summarized, thusly,

[http://[NMAH/]ark:/NAAN/Name[Qualifier]

An ARK is represented by a sequence of characters that contains the label, "ark:” optionally preceded by the protocol name ("http://") and hostname that begins every URL. That first part of the URL, or the "Name Mapping Authority Hostport" (NMAH), is changeable and replaceable, as neither the web server itself nor the current web protocols are expected to last longer than the identified objects. This part makes an ARK into an actionable URL, *i.e.*, clickable in a web browser. The immutable, globally unique identifier follows the "ark:" label. This includes a "Name Assigning Authority Number" (NAAN) identifying the naming organization, followed by the Name that it assigns to the object. Semantic opaqueness by using numbers in the Name is highly recommended. The Qualifier is optional and may be used to support access to variants of an object (different versions, languages, formats).

A list of the Name Mapping Authority is located in a file which is updated on an ongoing basis and is available for copying over the internet from the California Digital Library (CDL, 2009). The file contains comment lines explaining the format and giving the modification time, reloading address, and NAA registration instructions. (Kunze and Rodgers, 2008).

The following three ARKs taken from Kunze and Rodgers (2008) are synonyms for one object and the ark:/NAAN/Name remains the same.

http://loc.gov/ark:/12025/654xz321
http://rutgers.edu/ ark:/12025/654xz321
ark:/12025/654xz321

The NAAN part, following the "ark:" label, uniquely identifies the organization that assigned the Name part. Often the initial access provider (the first NMAH) coincides with the original naming authority (represented by the NAAN). The NAAN designates a top-level ARK namespace and once registered for a namespace it is never reregistered (Kunze and Rodgers, 2008). However, access may be provided by one or more different entities instead of or in addition to the original naming authority. The ARK resolution model is depicted in **Figure 5**.

The ARK concept is not commercially driven and has fairly low technical requirements (DNS, Web server and a Web browser on client side). The cost of participating is low (*e.g.*, no subscription fee). Any institution can obtain a NAAN by contacting CDL. Once it has an NAAN an institution can begin generating ARKs by using any software that produces identifiers that conform to the ARK specification. CDL uses an open-source application called “noid” (nice opaque identifiers). See Kunze (2005) for a detailed discussion of the noid application. As of January 8, 2009, thirty organizations had registered for Name Assigning Authority Numbers (NAAN; CDL, 2009).

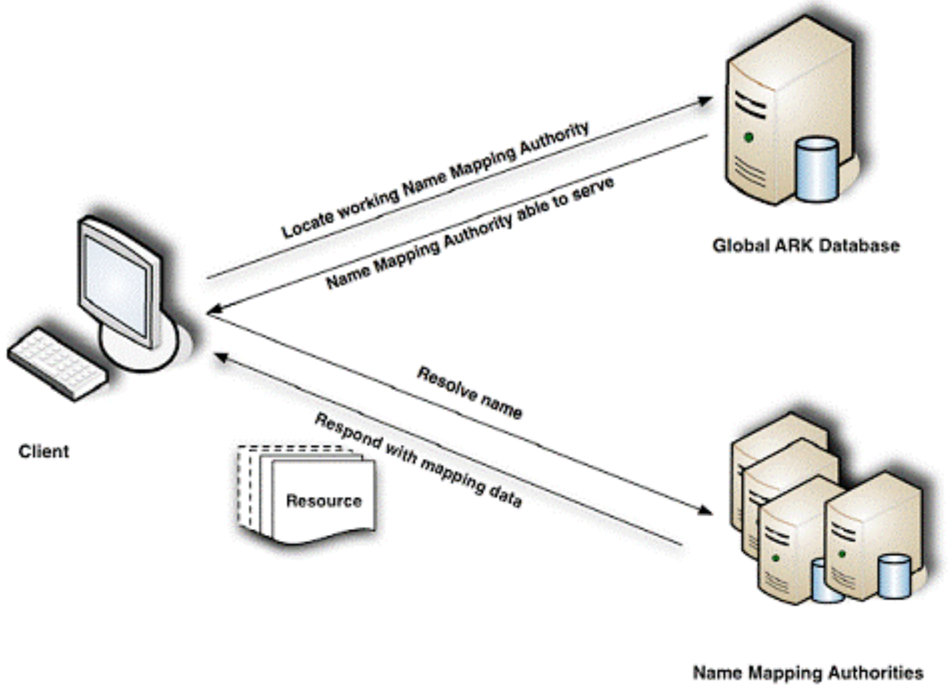
Locate working Name Mapping Authority
Name Mapping Authority able to serve
Global ARK Database
Client
Resolve name
Respond with mapping data
Resource
Name Mapping Authorities

Figure 5 ARK Resolution: The client sends the ARK identifier (format http;//NMAH:ark:NAA:Name) to the Global Ark Database to find a working NMA. The Global Ark database responds by sending the client the NMA able to serve the request and then redirects it to the appropriate NMA to resolve the name. The NMA responds with the mapping data to retrieve the requested object.

Notable Design Features of the ARK scheme and comments

- The scheme is standards based, protocol/technology independent and designed to meet the needs of digital archivists (Paradigm, 2008).
- ARKs can be used to identify different types of objects, such as agents and events as well as digital archival objects and metadata records (Paradigm, 2008).
- ARKs can be used in either a restricted or an open-access environment.
- The ARK system has written into its requirements the importance of organizational commitment to the ARK scheme (Paradigm, 2008).
- It is maintained by CDL, a leader in the field of digital preservation.
- The participation model used in the ARK scheme is more flexible than some of the other PID schemes: if one institution acts as both NMAH and NAA, it has complete control over its own identification scheme; since several NAAs can be connected to one NMAH, it is possible for one institution to host the digital archives of other institutions (Paradigm, 2008).

- Participating institutions can have an impact on the development of the ARK scheme since it is still a work in progress (Paradigm, 2008).
- Since the ARK scheme is fairly new (2001), it is difficult to assess how widely it will be implemented.
- "Some elements of the scheme [may be] superfluous to the requirements of digital archives, *e.g.,* hierarchies and variants can be defined using METS and PREMIS metadata rather than complex identifiers. In reality, it is probably more straightforward to use a simple single-level sequence of identifiers" (Paradigm, 2008).
- Given that XML is becoming the standard for encoding metadata, the current use of Electronic Resource Citation (ERC) for recording ARK metadata "may involve both duplication of metadata and the additional task of converting that data into a format that is not likely to be used elsewhere" (Paradigm, 2008).

Life Science Identifiers (LSID)

The Life Sciences Identifier (LSID) is a relatively new naming standard and data-access protocol developed in 2003 through the collaborative efforts of the (now defunct) Interoperable Informatics Infrastructure Consortium (I3C), IBM and other technology organizations such as Oracle, Sun Microsystems, and the Massachusetts Institute of Technology. The concept was to develop a common standard for data retrieval via the web so that scientists and researchers around the world would be able to share data, thereby facilitating collaborative efforts on projects such as drug discovery and other disease research (Salamone, 2002). The LSID and LSID Resolution System (LSRS) (Clark, 2004) were designed to provide simple solutions to the problem of identifying locally named objects that may be widely distributed over the network and across multiple knowledge bases. A client application resolves an LSID against a special server called an authority to discover data and information about the data (metadata) (Hobern, 2004; Clark, 2004; Object Management Group, 2003).

The LSID scheme encompasses the LSID and LSRS. Life Science Identifiers (LSIDs) are persistent, location-independent, object identifiers for uniquely naming biologically significant objects including but not limited to individual genes or proteins, or data objects that encode information about them. LSIDs are intended to be semantically opaque, that is to say they do not describe the characteristics or attributes of the object to which the LSID refers. LSIDs are expressed as a URN namespace and share the following functional capabilities of URNs (OMG, 2004).

The syntax of a LSID is as follows:

URN: Protocol:<AuthorityID>:<AuthorityNamespaceID>:<ObjectID>[:<Version>]

Examples of LSIDs that were taken from Clark (2003)

URN:LSID:ebi.ac.uk:SWISSPROT.accession:P34355:3
URN:LSID:rcsb.org:PDB:1D4X:22
URN:LSID:ncbi.nlm.nih.gov:GenBank.accession:NT_001063:2

Each LSID is prefixed by "urn" indicating that the LSID is a Uniform Resource name (URN), "lsid" indicates that the identifier is resolved using the LSID protocol, then follow the authority, namespace, and an object id sometimes followed by an optional revision component to indicate the version of the object. The authority is a domain name that can be resolved by the Internet DNS (typically a domain name owned by the data provider) and the namespace and identifier are specific to the data source which provides the object.

Note that the uniqueness of the LSID is in part guaranteed by the use of Internet domain names, which are globally unique. Providing that the data source ensures that each combination of namespace and identifier is unique within that data source, the LSID itself will be a globally unique identifier. Given a LSID, client software can retrieve metadata and/or data identified by that LSID.

LSID Resolution Services exist for life sciences data sources including GenBank, PubMed, Swiss-Prot, GeneOntology, and ENSEMBL (Szekely, 2003). In 2008 the Catalogue of Life included in its Annual Checklist LSIDs combined with LSID resolution service to provide a persistent and location independent means to access taxon metadata. It is difficult to ascertain the actual number of organizations which have implemented LSIDs..Szekely (2006) suggest to “just Google urn:lsid’ to discover its adopters.

Notable Design Features of the LSID and comments

- LSIDS allow local and shared management and have no upfront registration costs (Hagedorn, 2006).
- LSIDs allow for a distributed control of globally unique identifiers. Institutions have the option to register namespaces with a central Global Bidodiversity Information Facility (GBIF) authority rather than set up their own LSID authority (Hagedorn, 2006).
- The issue of truly persistent identifiers is separated from the management of the conventional URLs (which are both semantically and management wise overloaded, causing great instability) with LSIDs (Hagedorn, 2006).
- The convention that resolves a LSID returns metadata in Resource Description Framework (RDF). This has the potential of facilitating information integration from multiple sources using tools being developed for the Semantic Web.
- The lack of the delivery of standard metadata that always points to the latest version of an object renders the versioning system less effective.
- LSIDs are not widely used among the biological databases because “core providers such as NCBI that provide stable identifiers and well-documented services have little incentive to add support for LSIDs” (Page, 2008)
- Unlike a URL, LSIDs do not currently have native browser support and thus require some form of client plugin or proxy web service to retrieve the metadata response from the resolver. There are several online resolvers including the TDWG proxy at http://lsid.tdwg.org/ or the Firefox browser with the LSID Plugin. Once installed, the Firefox plugin will allow LSIDs (with the "lsidres:" prefix) to be copied into the browser's address bar (Catalogue of Life, 2008)

- The current mechanism for resolving LSIDs is not supported by existing Semantic Web tools (Page, 2008).

Handle System (Handle)

The Handle System technology was initially developed with support from the Defense Advanced Research Projects Agency (DARPA) by the Corporation for National Research Initiatives (CNRI) which continues to develop and manage it. The framework for this system was developed by Kahn and Wilensky in 1995.

The Handle System, which is logically centralized, physically distributed and highly scalable, includes an open set of protocols, a namespace, and a reference implementation of the protocols. The protocols enable a distributed computer system to store names, know as handles, of arbitrary objects and resolve those handles into the information necessary to locate, access, and otherwise make use of the objects. This information can be changed as needed to reflect the current state of the identified object without changing its identifier, thus allowing the name of the item to persist over changes of location and other related state information[5]. Each handle may have its own administrator(s) and administration can be done in a distributed environment. The name-to-value bindings may also be secured, allowing handles to be used in trust management applications (Lannom, 2000; Handle, 2009).

The Handle System, which has been designed from the start to serve as a general-purpose naming service, is also designed to accommodate very large numbers of entities and to allow distributed administration over the public Internet. It is hierarchical with the top level consisting of Handle.net services including the Global Handle Registry (GHR) which is operated by CNRI, which includes the Handle System Advisory Committee, and the lower level consisting of all the local Handle services (Sun *et al.,* 2003). An organization or individual (referred to as a Resolution Service Provider) that wants to provide identifier and/or resolution services using the Handle System technology obtains a handle prefix value from the GHR and becomes known as a Naming Authority (since it has authority over its own local namespace). The GHR registers, maintains, and resolves the Naming Authorities of locally-maintained Handle Servers. Any local Handle Server can, therefore, resolve any Handle through the Global Resolver.

The Handle system provides a resolution system optimized for speed, reliability, scaling with the focus on identifying objects, not servers. There is separation of control of the handle and who runs the servers; distributed administration, with granularity at the handle level; all transactions can be secure and certified.

The interoperable network of distributed Handle resolver servers are linked through a Global Resolver (http://hdl.handle.net/). When the browser receives a handle, it resolves it through a local handle server; if it cannot recognize the local resolver, the handle is sent to the global registration service for resolution and redirection to the appropriate local handle server which, in turn, can resolve the local part of the identifier to the object in a resource server (repository). Handles, as most commonly used, resolve to the current URL of an object. The HANDLE resolution model is depicted in **Figure 6**.

[5] In this context “State information” refers to information about a specific object

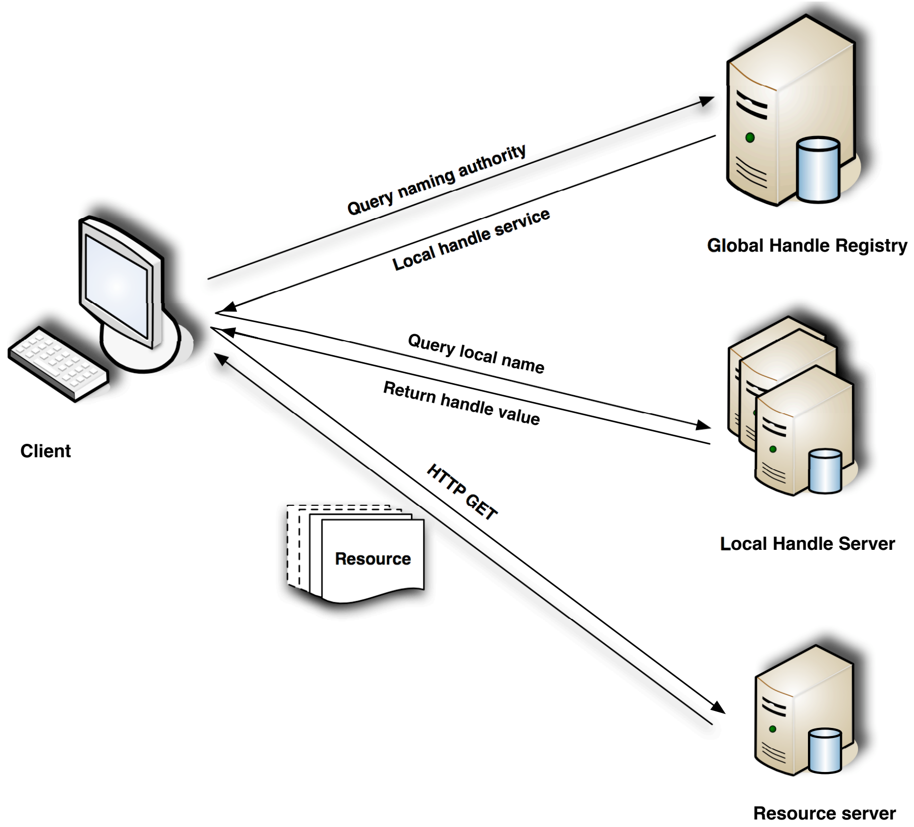

Figure 6: HANDLE Resolution. HANDLE identifiers are passed by a client as a query of the naming authority to the Global Handle Registry (GHR) using the Global Resolver http://hdl.handle.net. The GHR responds by sending the client the address of the local HANDLE server; a query is then sent to the local HANDLE server which returns the HANDLE value that is then used to form an HTTP GET to retrieve the object of interest from the web server hosting the requested version of that content.

The syntax of a HANDLE consists of two parts:

[Naming Authority]/[LocalName]

The Handle must be a string composed of UTF-8 characters beginning with a Naming Authority followed by a slash and then a Local Name representing an object identifier (Sun, *et al.,* 2003). No semantics are necessary in the identifier string

Examples of a HANDLE are:

10.1099/ijs.0.64483-0
2027/spo.3336451.0010.202

The Global Handle Registry assures that each Naming Authority is unique; local Handle Servers assure that each object identifier assigned by a naming authority is unique within that naming authority. The object identifier portion of a Handle can be an intelligent string or an unintelligent "dumb string" because within the Handle system, all identifiers are dumb strings. Many organizations use identifiers already developed for their internal systems in this portion of the Handle (the syntax allows unlimited numbers of Unicode characters, so there will be no problems in translating an existing string).

The Handle System also supports the simultaneous return as output of multiple pieces of current information related to the object, in defined data structures (multiple resolution), and priorities can be established for the order in which the multiple resolutions will be used. Handles can, therefore, resolve to different digital versions of the same content, to mirror sites, or to different business models (pay versus free, secure *vs*. open, public *vs*. private). They can also resolve to different digital versions of differing content, such as a mix of objects required for a distance-learning course. For example, one Handle could provide the capability to access all of the digital materials for a course (this would of course require additional appropriate definition of structured metadata). In addition to URLs, Handles can resolve to email accounts or to other Handles (supporting various Web services applications). Each of these various target categories has a unique data type, and the list of types can be extended. Because current Web browsers do not support the Handle resolution directly, it is necessary to use intervening software that can be downloaded as an add-on client or hosted on a proxy server.

The Handle System is currently being used by a number of different institutions and projects. As of May 2008, there were 3,797 Handle Naming Authorities and over thirty five million Handles (Lannom, 2008).

Notable Design Features of the Handle System and comments

- As one of the first Persistent Identifiers schemes Handle has been widely adopted by public and private institutions and proven over several years. Major publishers use Handle for persistent identification of commercially traded and Open Access content through its implementation with the DOI (see below). It is maintained by a well-known international organization with a commitment to ensure its continuation, so it is stable and well-established (Paradigm, 2008).
- The Handle system is well-documented (www.handle.net)
- It conforms to the functional requirements of the URI and URN concepts, and is independent from, yet interoperable with, current protocols like HTTP (Paradigm, 2008).
- Handles can be resolved natively in browsers (*e.g.*, Internet Explorer and Firefox) by means of an extension (see Handle System Client Extension, 2009) or the Handle Proxy Server.
- Multiple resolution and well-defined extensible data typing are available for the construction and definition of applications beyond simple one-to-one resolution.
- The Handle system makes explicit the importance of organizational commitment to a persistent identifier scheme.
- Handle syntax is straightforward and existing local identifier systems can be easily incorporated into Handles if required (Paradigm, 2008).

- Handle allows for different levels of access required for managing personal digital archives: operations on the Handle registry are tightly controlled by a detailed authorization mechanism to ensure data security; and Local Handle Servers can be configured to allow either internal or external access (Paradigm, 2008). Administration can be specified at any desired level of granularity down to individual Handle records (in contrast to DNS-based systems which require a hierarchical top-level domain name administrator).
- Since it is a distributed model, local Handle services and Naming Authorities have the autonomy to manage their own Handles (Paradigm, 2008).
- Institutions can share a local service under the same Naming Authority since Handle is scaleable.
- Although the resolution of Handles is free, there is a one-time Registration fee (currently $50 per year, with options for larger scale implementations at volume discount) for new prefix assignment to help defray the cost of running the GHR and an Annual Service Fee for Resolution Service Providers. This is similar to the cost of Domain Name registration.
- While there are authorization mechanisms, current public implementations have a strong emphasis on identifying objects which are openly available via the Web, rather than held in the more restricted context of a digital archive. With the increased proliferation of such objects on the web, *e.g.*, even pay-for-view is available on the Web this becomes less of an issue (Paradigm, 2008).
- Persistence is not necessarily required nor is an appropriate social structure provided (Garrity, 2006) although it is expected that major uses of the system will wish to use Handles as persistent identifiers and develop such social infrastructures in applications, such as the DOI System.
- Given that the character set for Handles is much broader than for URIs, institutional naming policies have to place restrictions on the characters used in order to comply with URI requirements (Paradigm, 2008).
- Handle server must be installed and managed by local technical staff and there is no ongoing technical support (Garrity, 2006).
- The Handle System provides a well-managed resolution component but agnostic as to accompanying metadata schemes

Digital Object Identifier System (DOI)

The creation of the Digital Object Identifier (DOI) system was initiated by the Association of American publishers in order to help the publishing community manage digital rights (copyright compliance) and electronic commerce. The DOI System provides an infrastructure for persistent unique identification of entities, designed to work over the internet. A DOI name is permanently assigned to an object, to provide a persistent link to current information about that object, including where the object, or information about it, can be found on the internet. While information about an object can change over time, its DOI name will not change. Although the system arose from an initiative of the publishing industry, and its focus is on objects of information content it is expanding into related areas (*e.g.* data, parties, licenses). DOIs can potentially be applied to all forms of content (*e.g.*, articles, books, images, data, files, legal documents, etc).

The DOI system provides resolvable, persistent, semantically interoperable identification of objects in a networked environment, and so enables the construction of automated services and transactions. Applications of the DOI System include but are not limited to: managing content location and access; managing metadata; facilitating electronic transactions; persistent unique identification of any form of any data; commercial or non-commercial transactions.

A DOI name can, within the DOI system, be resolved to values of one or more types of data relating to the object identified by that DOI name, such as a URL, an e-mail address, other identifiers, and descriptive metadata. The DOI System resolution component is an implementation of the Handle System. DOI adds additional technical and social infrastructure in order to provide a full application service for identifiers

The content of an object associated with a DOI name is described unambiguously by DOI metadata, based on a structured extensible data model that enables the object to be associated with metadata of any desired degree of precision and granularity to support description and services. The data model supports interoperability between DOI applications.

The DOI system, which is managed by the International DOI Foundation (IDF founded in 1998 as an open membership consortium including both private and public partners), is currently a draft International Organization for Standardization (ISO). This system consists of several existing standards-based components, most notably the Handle resolution system and the indecs Data Dictionary (INDECS)[6] which have been brought together and further developed to provide a consistent system (Paskin, 2005a). As a result DOIs can be used for management of data in both commercial and non-commercial settings.

The DOI system is composed of a numbering syntax, a resolution service, a data model, and procedures for the implementation of DOIs. Any existing numbering scheme and any existing metadata scheme that provide an accepted numbering or descriptive syntax for a particular community or area of interest (such as formal ISO standards or accepted community practice) can be used within the DOI System (IDF, 2008). A DOI may be assigned to any item of intellectual property, which must be precisely defined by means of structured metadata. The DOI itself remains persistent through ownership changes and unaltered once assigned.

The syntax of a DOI consists of two parts:

10. Prefix/Suffix

[6] INDECS encompasses a generic metadata analysis, a high-level metadata dictionary, principles for mappings to other schemas. It uses a sophisticated model to identify and describe intellectual property items from data sources previously considered incompatible…such as the copyright societies' CIS, the recording industry's DCMS, the library community's FRBR, the museum/archive community's CIDOC reference model, and the book industry's EPICS/ONIX. For more information go to http://cordis.europa.eu/econtent/mmrcs/indecs.htm

The prefix denotes the naming authority and always begins with 10. and is followed by the number assigned to a specific registrant by a DOI Registration Agency (RA). DOI Registration Agency status may be any profit-making or non-profit-making organization that can represent a defined "community of interest" for allocating prefixes to registrants. A fee is paid by RAs to the IDF in recognition of their participation in, and their ability to build a business using, the DOI System (IDF, 2006).

Registrants can be any individual or organization that wishes to uniquely identify intellectual property entities using the DOI System. The suffix is assigned by the Registrant to identify a single object. A DOI is represented using the Unicode character set and is encoded in UTF-8 (Paskin *et al.*, 2003).

Three examples of DOIs are:

10.1000/182
10.1038/nmeth1156
10.1093/bioinformatics/bti346

In the above examples these three different prefixes may belong to three different Registrants or they may belong to one Registrant. The suffixes assigned by the Registrant to specific content can follow any system chosen by the Registrant, and be assigned to objects of any size – a book, an article, an abstract, a figure, a legal document -- or any file type -- text, audio, video, image or an executable. An object may have one DOI, and a component within the object may have another DOI. The suffix can be as simple as a sequential number, or a combination of numbers and letters.

Because digital content may change ownership or location over the course of its useful life, the DOI system uses a central directory (Paskin, 1999). A DOI can be resolved into a URL using this resolution service, such as the DOI System Proxy Server, or the Handle System Proxy. The DOI resolution model is depicted in **Figure 6**. When a user clicks on a DOI, a message is sent to the central directory where the current web address associated with that DOI appears. This location is sent back to the user's Internet browser with a special message telling the system to "go to this particular Internet address." The user sees a "response screen" -- a Web page -- on which the requested object itself appears or, if not, then further information about the object, and information on how to obtain it. When the object is moved to a new server or it is sold to another company, one change is recorded in the central directory and all subsequent users will be sent to the new site. The DOI remains reliable and accurate because the link to the associated information or source of the content is so easily and efficiently changed. The underlying technology used in the DOI system is optimized for speed, efficiency, and persistence.

Initial applications of the system are simple redirection to URL; more sophisticated functionality is available, through multiple resolution, data typing, and Application Profiles based on structured metadata, and these are expected to proliferate. The DOI System is currently a Draft ISO standard.

At the end of 2008 there were eight DOI Registration Agencies located around the globe (Lannom, 2008). As of January 08, 2009 one of these agencies alone, Crossref, which has 2,684 participating publishers and societies, had registered 34,929,227 (Crossref, 2009)

DOIs are being adopted for use to keep track of materials by wide variety of organizations, including the European Commission (EC), and the Organization for Economic Co-operation and Development (OECD) (Paskin, 2005). DOIs are being used by repositories to identify data, *i.e.*, to provide entry points to scientific data that cannot be categorized as “documents”, such as scientific measurement data or similar information. For example, the project of The German National Library of Science and Technology (TIB) uses DOIs to persistently identify scientific data set; for more information see Paskin (2005) and STD-DOI: Publication and citation of Scientific Primary Data (2008). Experimental DOIs are also being for tracking and managing dynamic terminologies, such as biological names of organisms, genes and gene products that provide crosslinks to other digital content that is specifically related to a name or term, in temporal context (N4L, NamesforLife).

Notable Design Features of the DOI System and comments

- The IDF provides a full identifier system, with technical and social infrastructure to enable applications; social infrastructure guarantees persistence. Its policies ensure DOIs continue even when RAs fail. The IDF and member RAs are persistent as it is self-funding (IDF, 2006; Garity, 2006).
- DOI names can be resolved natively by adopting one of a range of available appropriate tools (http://www.doi.org/tools.html), or via http using a standard proxy server (http://www.dx.doi.org).
- It adopts the proven Handle System technology and adds specific resilience and performance improvements for the DOI application (*e.g.*, improved database handling, separate proxy server distribution).
- It provides a proven data model which can accommodate existing metadata schemes and enable interoperability with other schemes.
- It provides an infrastructure for implementing a comprehensive digital identifier system, while still allowing each RA a considerable degree of autonomy to implement their own system and business model *e.g.*, there is a scope for establishing an RA for those working with biological digital archives (Paradigm, 2008).
- The option to create a 'Restricted' Application Profile means that the scheme could be used in a non-public digital repository environment as well as an open environment (Paradigm, 2008).
- Interoperability has been maximized by being standards based.
- Existing local identifier schemes can be easily incorporated into DOI System if required.
- There are upfront costs (*e.g.*, entry fee, annual fees, *etc.*) associated with the implementation of a DOI scheme in return for participation in an existing scheme. The seed funding (in the form of a loan) that was provided to the IDF was in the seven figure range. There are annual dues of $11,500 for general members and $40,000 for Registration Agencies to be part of the DOI community. This provides the operating funds to sustain the technical and social infrastructure.

- A high level of commitment from its participating members with strict guidelines as to its usage is required.

Some Users of Persistent Identifiers

According to the National Collaborative Research Infrastructure Strategy (NCRIS) modern research is increasingly dependent on technological platforms that enhance the ability of the research community to generate, collect, share, analyze, store and retrieve information. Ideally, such information will exist within an environment of enabling and value-adding information services that support object location, object access, and object analysis (Ward and Macnamara, 2007). The internet and the web provide the platform upon which such services can be offered.

When first introduced, the internet was a medium mainly for scholarly communication and remote control of computers, but with the global acceptance and use of the web it has become a standard medium for publications and access to all types of information. It has evolved into a global mechanism for mediating an endless variety of commercial and non-commercial transactions involving tangible and intangible assets.

Today commercial publishers and numerous academic and cultural institutions offer a wide variety of objects on the Web. Doctoral students publish their theses electronically, and in digitization projects huge amounts of paper materials are converted to electronic documents, such as the previously mentioned Biodiversity Heritage Library. The advantages of electronic publication for instant access and easy, low cost distribution and duplication are obvious (Hilse and Kothe, 2006). But it is not just about scholarly content or electronic publications. It is about virtually anything that can be traded and tracked. What we are interested in are tangible assets that have extensive metadata and associated linked information (*e.g.*, organisms, genomes, genes, content, rights and obligations, contracts, *etc*.). The identifiers simply provide a means of identifying all things of interest through well defined metadata and providing a method to persistently link together the related objects, revealing not just past or current relationships, but also relationships that will emerge in the future (Garrity and Lyons, 2003).

At the same time, many internet users believe that most things on the Internet (especially text) should be free (Davidson, 1998), thereby making the internet a difficult environment for commercial publishing, as was demonstrated when the popular online magazine *Slate* initiated a $20 annual subscription rate year. Overnight, readership plummeted from nearly 60,000 to a paid subscription list of about 17,000 (Pogrebin, 1998). Despite this somewhat hostile climate, publishers of the most costly scholarly journals, mainly those produced by the scientific, technical, and medical publishers, realize that long-term survival depends on their ability to market products successfully over the Internet. In the fast-changing world of electronic publishing, there is the added problem that ownership of information changes and location of electronic files changes frequently over the life of a work. Within the past five years the majority of the Scientific, Technical, Medical and Scholarly (STM) publishers have established a significant Web presence. John Wiley & Sons, Springer, Elsevier Science, and many other publishers have duplicated their print journal output in electronic format, and made them available on the Internet. Several publishing journals are now only available electronically. As a result of this industry wide shift from paper format to global digital format, the publishing industry has been a primary driver in the development and implementation of persistent identifier schemas. Crossref was created as an independent membership association by STM publishers in order to connect users to primary research literature through a DOI RA that performs reference cross-linking, subject to publisher-access controls (Garrity, 2007). Government, libraries and a growing number of organization managing large digital databases are already endorsing and using DOIs. The only other identifiers used widely in the life sciences (and specifically the health-related sciences) are PubMed IDs (PMID) for content and GenBank identifiers (INSDC) for gene sequences. The issues faced by electronic publishers holds true for all industries and organizations that use the Internet to provide products and services.

It should be noted that an organization may choose to use one or more of these persistent identifier schemes to meet its need. Within the U.S. Government, for example, the Government Printing Office (GPO) and the Office of Scientific and Technical Information (OSTI) of the Department of Energy (Energy) use PURLs and their own installations of the PURL Resolver to manage their connections to the full text of documents. The Handle System is used by The Defense Technical Information Center (DTIC) to control the identification and location of objects it receives from throughout the Department of Defense. DTIC is a Handle Naming Authority. Additionally, DTIC is exploring ways to make a variety of digital materials available to its user communities through a Defense Virtual Information Architecture (DVIA). Materials to be included range from textual materials such as technical reports and electronic journals, to videos, photographs, audio recordings, maps, and possible medical imagery. Since DTIC plans to store some of the materials in its own repository and provide links to remote sites when linking is the best way to deliver the information it will also include extensive searching capabilities. The Handle System is considered to be an essential component of this application (CENDI, 2004).

These are just a few of the institutions that have integrated the use of persistent identifier schemas into their digital platforms. Before doing so, they each had to address a number of issues to determine which persistent identifier scheme would best fit into their digital management structure. Hilse and Kothe (2006) and Paskin (2008b) recommend strongly that organizations collaborate with partners that have existing schemes and similar problems to solve and to choose the syntax for their persistent identifiers in such a way that they can be integrated into any of the schemes introduced in this report.

Issues

The sheer number of digital assets being produced and stored by increasing numbers of organizations has made clear the need to better manage, locate and retrieve these objects over time. Obviously data management, of which digital preservation is a part, does not just happen. It must be carefully planned and well-implemented. This recognition has led to the increasing adoption of data management plans (DMP). Choosing the persistent identifier scheme best suited for the needs of the institution is an important part of any DMP. It should play an important role in the development of an International Regime as part of the ABS of the CBD. The use of such an identifier would greatly facilitate the monitoring and tracking of the use of genetic resources. In fact, during the January 2007 meeting of technical experts on an internationally recognized certificate of origin/source/legal provenance the group recommended that persistent identifiers should be used and that an international registry containing the identifier of the certificate could serve as a clearing house mechanism (Ad Hoc, 2007). The Australian Department of the Environment and Heritage and Australian National Competent Authority on Genetic Resources has already set up a system through which one can apply on line for virtual permits for access to genetic resources for commercial or for non-commercial purposes for access (ABR, 2009)

Prior to the implementation of any persistent identification scheme, there are a number of concerns that need to be carefully and thoroughly addressed (Paskin, 2008a, 2008b; Davidson 2006; Broeder, 2007; Cendi, 2004; Hobern, 2004; Bellini, 2008; Garrity, 2007). Some questions that should be answered include:

- What will the identifier be identifying — the object, an abstract representation, or a physical object with associated metadata? How will the referent (the object which is identified) be precisely defined in such a way as to be understood by other users outside the control of the assigner? What metadata scheme will be used to do so?
- What will the identifier be required to resolve to: location, metadata, services? How can we avoid conflating “referent of the identifier” with “what the identifier resolves to” (not necessarily the same thing at all - though that may be intended!) – this conflation often arises due to the case with URL referencing.
- Does the identifier need to be globally or locally unique?
- What level of granularity is needed and will opaque or semantic identifiers be assigned?
- Are there legacy naming systems that need to be incorporated? If so, how will interoperability between naming systems be handled?
- At what point does an object change enough that it requires its own identifier?
- How will metadata be stored and bound to the identified object?
- Will the identification scheme of today be able to meet future needs?
- When is an identifier applied to an object and who will manage the identifiers over time?
- How will the assignment and long-term management of identifiers be financed?

Global *vs* local uniqueness

It is often said that a certain class of identifiers must be ‘‘globally unique.’’ That is, they can be used anywhere in any system and will never overlap with an identifier assigned by someone else. This becomes an increasingly important concern with the growing number of interactions among different systems in the digital and networked world. The common experience is that an identifier is created within a system or within a given context thereby being locally unique but not necessarily globally unique should the object be accessed at a later date in another or larger context.

Persistence

Persistence refers to the permanent lifetime of an identifier. It is not possible to reassign a persistent identifier to other objects or to delete it (other than for valid reasons such as error corrections). That is, the persistent identifier will be unique (within the context in which it was assigned) forever, and may well be used as a reference to that object far beyond the lifetime of the identified object or the naming authority involved. The only guarantee of the usefulness and persistence of identifier systems is the commitment shown by the organizations that assign, manage, and resolve the identifiers. As Paskin (2008a) notes, persistence is not a technical issue, it is a social issue. Persistence can also be seen as a form of interoperability requirement (it is "interoperability with the future").

Resolvability

Resolvability refers to the possibility of retrieving an object (or information about it) if it is available on the web or some succeeding network in the future. It is important to distinguish the concept of identification from resolution. The choice of the identification namespace does not necessarily imply choosing corresponding resolution architecture.

Governance

To assure reliability of a persistent identifier scheme, two aspects have to be assessed: its infrastructure must always be active (service redundancy, back-up deposit services, *etc.*) and the register updated (through automatic systems). The only guarantee of the usefulness and persistence of identifier systems is the commitment shown by the organizations that assign, manage and resolve the identifiers. For example, in the cultural heritage domain the tendency is to make use of services provided by public institutions like national libraries, and state archives (Bellini, 2008). Requirements such as the authority and credibility of the organizations offering such services should be carefully evaluated before adopting a solution. Given that ABS is a global issue, it is critical that the chosen persistent identifier system be one which has a strong, durable infrastructure and that the administrative organization is trusted, well-respected and enduring.

The ARK and PURL specifications describe systems that can be hosted by almost institution. The Handle server can also be locally installed in a manner similar to any other Web server. By comparison, the DOI identifier is centrally administered by the IDF and has a number of associated fees. However, as Tonkins (2008) observes fees are often associated with reliability, authority, longevity, and durability.

Granularity

An identifier system will be more effective if it is able to accommodate the special requirements of different types of objects that are made available in digital form. An identifier system should be able to manage different levels of granularity because what an "identifier" must point to can differ considerably in the different user application fields. Granularity refers to the level of detail at which persistent identifiers will need to be assigned. The granularity requirement has considerable impact on the identifier system an institution adopts. In some situations, it may be necessary to cite a web page that serves as access to a collection of web files, or to cite a journal article, an item, or a chapter. However, because of rights management, some finer details may be required. Each institution should evaluate whether a persistent identifier scheme provides the right level of granularity for their type of objects.

In practical terms the process of monitoring ABS compliance is fundamentally an issue of rights management. In the selection of an identifier for this purpose, it should be noted that the greater the degree of its granularity, the larger the pool of potential benefits that can flow from its implementation and use. For example, with a finer level of granularity, genetic resource providers could even derive some benefits from the non-commercial use of their biological and genetic resources in various types of products, such as a "modeling fee" for the images of their flora and fauna used in the glossy publications of large museums, and various wild-life organizations. Royalty payments derived from reuse in other publications and products (*e.g.*, calendars, advertisements, commercial publications) could also be identified and tracked to ensure compliance and instances of piracy. With respect to the identifiers currently under review, DOIs are the only ones that were developed with rights management as a critical part of its underpinnings.

Interoperability

Interoperability--the ability of different systems to exchange information and to use the exchanged information--is fundamental for guaranteeing the possibility of diffusing and accessing objects. An object can be part of more than one domain, and can be identified by different systems; as noted previously, an organization may adopt several different schemes for valid reasons; so it is necessary to guarantee interoperability among different identification systems as well as implementations based on the same namespace. Many technologies and approaches are available and some of them are tailored for specific sector requirements. Among different systems interoperability must be realized at least at the service level offering common and easy user interfaces. System interoperability can be based on the adoption of open standards. Three sorts of interoperability can be distinguished:

- Syntactic interoperability. The ability of systems to process a syntax string and recognize it (and initiate actions) as an identifier even if more than one such syntax occurs in the systems.
- Semantic interoperability. The ability of systems to determine if two identifiers denote precisely the same referent; and if not, how the two referents are related.
- Community interoperability. The ability of systems to collaborate and communicate using identifiers whilst respecting any rights and restrictions on usage of data associated with those identifiers in the systems (Paskin, 2008b; IDF, 2008).

Opacity

A persistent identifier should not contain any information about the object it identifies (opaque id); rather it should consist of random characters/numbers that have no associated semantics. Opaque strings prevent any possible misunderstanding or poor translation of the semantics of an identifier. Opaque identifiers can be chosen by automated means using NOID (nice opaque identifier) or UUID/GUID (universally unique identifier).

In most cases, when a decision is made to use non-opaque identifiers, names are deliberately chosen to assert fact (Kunze, 2008). It is generally easier for a person to memorize and use mnemonic-based identifiers rather than those that contain a meaningless character sequence. However this has no relevance to machine processing. "There is a trade-off between the ability to track down an object, should the persistent identifier fail to resolve, from the semantic information available in the string, and the increased likelihood that a string containing meaningful semantics will at some point be altered" (Tonkin, 2008).

Metadata

Persistent identifiers allow access to objects as well as to their associated metadata (but only in those systems in which metadata is a part of the identifier specification and implementation), which is fundamental for enabling users to identify content. Therefore, with the development of global databases and technological innovations it is evermore important to develop advanced metadata management and user services, such as services that extends to different repositories (Bellini, 2008). Metadata fields can also be used to manage the different versions of an object, each one requiring a separate persistent identifier. Further information about metadata can be had from *Understanding Metadata* by The National Information Standards Organization (NISO). Structured ontologies are the most appropriate form of defining interoperable metadata schemes to allow extensible knowledge representation (also foreseen as a requirement in Semantic Web activities) (Rust and Bide, 2000; Sowa, 2000).

Future internet architecture

Design of persistent identifiers should be cognizant of the fact that the original design of the internet emerged over 30 years ago, a time that predates both the personal computer and local area networks: further evolution is inevitable. A major influential study on future design noted that "it is possible to separate the ideas of location and identity, both of which are represented by the IP address in today's Internet, and that the resulting architecture facilitates mobility as well as solving other problems with today's network". (Clark *et al.*, 2003).

Discussion

Paskin (2008a) suggests that the successful implementation of persistent identifiers schemes within an organization may face more social and economic challenges than technical ones. A persistent identifier scheme requires ongoing maintenance and, therefore, ongoing resources. Allocation of resources is always a point of contention within an organization (Cendi, 2004). All schemes need indefinite support for at least a Web server, web browser, and domain names as well as indirection or redirection tables (Kunze, 2008). It is also work noting here that although it was not possible to determine the costs involved in developing each of the persistent identifier schemes it is known that the seed funding for the International DOI Foundation was in the seven figure range.

Sharing costs by funding a common identifier system tool may be a solution (*e.g*., more than 2600 publishers collaborate in assigning DOIs through CrossRef). At the individual agency level, the resolver must be kept up-to-date with the current URLs for the locations of the objects. This means that as usage grows so do the underlying costs of the necessary infrastructure. This cost that can become a significant factor when large numbers of identifiers are in use and mapping to URLs is an ongoing activity. The resolution provided by the system is only as up-to-date as the physical locations to which the persistent identifiers point. While some of this updating can be automated, responsibility for this updating and ensuring its reliability must be assigned within each agency, program or office or to a trusted third-party. In the case of integrating a persistent identifier scheme within the ABS process the use of a trusted third party is probably the best option. It is not sufficient to create identifiers and leave them without maintenance; active management is needed in order to gain the benefits of such a system. It is clear that active management is a key driver to an efficient and effective DMP.

Cultural norms and expectations inherent to the environment in which the organization functions need to be considered when investigating persistent identifier schemes. In the domain of biodiversity conservation and sustainable development, such as the CBD-ABS, trustworthiness and accuracy are critical. The persistent identifier scheme will be used to track and monitor the use of genetic resources in both commercial and non-commercial activities and must meet the needs and expectations of all the parties to different agreements (providers and users, government agencies, commercial concerns, *etc*.). This is especially true when these identifiers become a factor in compliance and intellectual property regimes.

The proper selection and implementation of a robust persistent identifier system will be critical for the enablement of many of the provision of the CBD. These same identifiers will become an integral part of an intertwined international biodiversity network that will be critical for many policy issues as well as enforcement and litigation. In the latter case, it will not be a case of “what if”, but "when". If the identifier system is robust and well designed it will play a critical role in ABS. If not, then ABS will not be easily or reliably achieved.

A summary of the identifier properties discussed above is given in **Table 2** below.

Property	**URN**	**PURL**	**ARK**	**LSID**	**HANDLE**	**DOI**
Global	**-**	**-**	+	**V**	+	+
Persistent	**-**	**-**	**?**	**?**	+	+
Resolvability	**-**	**V**	**V**	+	+	+
Governance	**-**	**-**	+	**-**	+	+
Granularity	+	**-**	+	**V**	+	+
Interoperability	**V**	+	+	**?**	+	+
Opaque	**-**	**-**	**V**	**-**	+	+
Metadata	**-**	+	**-**	**V**	+	+
Standards compliant	**W3C**	**-**	**-**	**-, +**	**-**	**ISO draft**

Table 2: A comparison of key properties of identifiers discussed in this report. + means supported; - means not supported; V means variable levels of support depending on implementation; ? means unspecified.

Applications using persistent identifiers

In 2006, the National Information Standards Organization (NISO) sponsored a workshop on identifiers (NISO, 2006). Perhaps one of the more interesting observations in the published report of the meeting was the realization that "Although used every day, identifiers are a mystery to many people, including people responsible for building complex information systems." In their simplest form, identifiers are synonymous with unique keys used in a database. When restricted to local use within a given data system (*e.g.,* a laboratory information system in a pharmaceutical or biotech company, a cataloging system in a library, an inventory control system in a warehouse), identifiers generally are not problematic because policies and procedures are in place to govern their creation and use. This ensures uniqueness and non-redundancy of the identifiers and allows their use in mission critical systems with a high degree of confidence.

Problems can arise when identifiers developed for use in one system are incorporated into other systems in which they are used as foreign keys. This is particularly problematic when there may be differences in the underlying schemas such that the mapping of data is not exact or where uncontrolled changes in identifier syntax and semantics may occur in either system. Problems also arise when identifiers intended for one application are subsequently applied to other applications that were neither intended nor anticipated; a phenomenon known as mission or feature creep. Problems also arise when identifiers are semantically laden that are either non-unique, difficult to model, independently managed, or all of the above, such as biological names (Garrity and Lyons, 2003).

As the life sciences have become increasingly data driven, identifiers of many different types have become quite familiar to contemporary practitioners. Among the most ubiquitous are INSDC identifiers (aka GenBank accession numbers), but many others also appear in the literature today. Each provides a potential means of linking to underlying data and other information sources. However, few provide any guarantee of persistence, uniqueness or actionability. Users are left to use those identifiers to query the data repositories of their choice by any number of methods.

Identifiers are recognized as a major source of difficulty in bioinformatics applications (Clark, 2003). Recently, resolvers and integration networks have been built to assist end-users in the mapping of various identifiers used in key resources (*e.g.*, the Genomic Rosetta Stone, EMBRACE Network of Excellence; SRS System); Van Brabant *et al.*, 2008). However end-users should understand that identifier mapping is only as current and reliable as the curatorial efforts expended by participating projects. Proper curation and maintenance of stable and enduring data systems is not an inexpensive activity, nor is it highly rewarded. This is particularly true in academic settings, where short-lived "bioinformatic resources" are often created as part of a thesis project using methods such as screen scraping and wholesale copying of data from public resources such as the INSDC repositories and biological resource centers without a clear awareness of the potential consequences, and then abandoned (Veretnik *et al.*, 2008). To the end-users of such temporary bioinformatic resources, we can only advise *Caveat emptor*.

This is not to say, however, that reliable and useful community resources cannot be created and maintained into the future. The technology is available and can be readily applied to provide a lasting solution to the problem of tracking genetic resources in such a way as to provide transparency to commercial and non-commercial users and providers. A system of such a design could also be developed that would provide a reasonable basis for equitable implementation of the ABS regime.

Persistent identifiers are a powerful enabling technology; for example, the use of DOIs for published articles allows rapid and accurate tracking of written works. When properly deployed, PIDs can provide direct access to specific information, at the point of need.

Those needs, however, differ from one end-user to the next and can range from access to laboratory data, articles in the scientific literature, or specimens of type or reference materials in a biological resource center to regulatory, patent or safety information about a particular organism appearing in the legal or gray literature. A confounding problem is that in any given system there is no guarantee that the terminologies and concepts used will be current or resolvable to those in current usage.

The manner and form in which the information is delivered to an end user (whether the end user is a human or another computer) can be specified in an application program interface (API) that allows developers to build tools that not only use the information that identifiers point to, but also define how such information is formatted and processed. This presupposes that the structure and contents contained in the objects to which PIDs point have been explicitly defined *a priori*. If so, then powerful applications can be built to meet the needs and expectations of a particular user community.

Perhaps the best way to provide the reader with an understanding of this potential is by example. One of the most widespread uses of PIDs is in the management of bibliographic information. ARK, PURL, HANDLE and DOI all find use in such settings where the chosen PID serves as a foreign key in a bibliographic database and points to the associated metadata for a specific publication. At present, the most widely used persistent identifier is the DOI. The DOI registration agency Crossref has assigned over 34 million DOIs to content produced by more than 2000 scholarly publishers (for-profit and not-for-profit, including open access content). When an end-user submits a DOI to the Crossref resolver (either directly using a plug-in extension in the Firefox browser or indirectly through the DOI or Handle proxy servers), they receive a response page that contains the front matter of the article and, depending on the end-user's rights, direct access to the content or to a "sales contract" that can provide access to the content on a pay-for-view basis.

Crossref registrants (typically publishers of STM literature) have access to tools that permit linkage in the bibliographic data of each of their published articles to previously published articles bearing Crossref DOIs. Crosslinking of bibliographies is a condition of Crossref membership. Newly published articles are also assigned a DOI so that each becomes integrated into the greater system. This provides a means of building a large and valuable network of interlinked content that can be followed over time with forward pointing links. The use of DOIs as opposed to URLs obviates the cost of having to maintain all of the links in previously published content. Maintenance of the DOI links and metadata is contractually guaranteed between the registrant and Crossref. DOI services for other types of content also exist, but most of those applications are not as well developed when compared to the services offered by Crossref. Among these are DOIs for books (Crossref and Bowker), and physical science data sets (TIB). Work is also underway to increase the granularity of information identified within a document such as assigning DOIs to tables or figures that could be used elsewhere. Such applications will provide a mechanism for ensuring that all rights and obligations associated with reuse are enforced.

As a globally unique persistent identifier (GUID) service, NamesforLife (N4L) was conceived as a way to disambiguate biological names and other dynamic terminologies in the life sciences and elsewhere. This method embeds N4L-DOIs directly into publishers' XML tagged content during the composition stage as a value-added service. The N4L

DOIs link to highly curated content in the form of extensively interlinked information objects that provide supplementary information about a particular name and all its known synonyms and homonyms, current, past and future taxonomic concepts, links to underlying data (*e.g*., genetic/genomic data, phenotypic data, safety), links to the original taxonomic proposals and subsequent emendations, information about the practitioners who defined each taxon, and information about the taxonomic methodologies used in the definitions and the relevant governing code of nomenclature. In addition to the scientific, technical and medical literature, N4L technology is directly applicable to other types of content (regulatory, patent, gray literature). NamesforLife has also developed a browser-based tool that provides similar capabilities for digital content that is available on the web. Although deployed as a DOI service with the obvious benefits of DOIs as a class of PID, the NamesforLife data architecture is identifier neutral, extensible, and able to work with multiple identifier types.

Two other emerging concepts that were anticipated by the practitioner information object in the NamesforLife data model (Garrity and Lyons patent application) are OpenID and ResearcherID (Bourne, 2008). These initiatives seek to use persistent identifiers as a means of identifying individual researchers/authors so that their associated metadata (current name, institution/employer, contact information, et.) can be readily available for use in a variety of applications to ensure correct linkage to their bibliometric data. A similar approach has been implemented by LinkStorm for some commercial applications.

CBD/ABS services

Each of the identifiers described above are used in well-defined settings in which both the data and metadata models of the underlying repositories were established *a priori*. The identifiers serve as a means of directly accessing a specific record or other form of digital content or the associated metadata. If the identifier is actionable, then it is possible to accomplish this using the familiar interface of a web-browser. However, with the use of web services that provide structured access to the content of interest automatically (*e.g*. from a database or application on a handheld device using embedded PIDs), similar results can be achieved where an interactive interface is not suitable. **Figure 7** provides a representation of how such system might be designed to provide such additional capabilities while fitting into an existing infrastructure.

Both the LIMS and the NamesforLife models teach the value of a central authority for registering specific events (as opposed to objects) according to a set of well-understood business rules. When such an authority is in place, it becomes possible to traverse a series of transactions backward and forward in time, even in instances where some ambiguity may exist. By drawing on highly interconnected information, it is possible to not only follow events, but to accurately recreate those events, when adequate documentation is available. Such a system would be highly advisable for monitoring the use of genetic resources, especially since there will be instances in which long periods of time may exist between the time PICs, MTAs, and Certificates of Origins (CoO) are executed and some commercial or non-commercial product results. Such a system would make it possible to enter the document trail for any genetic resource at any point in time and trace events backwards and forwards in time, including any rights or obligations that might be due one party or the other.

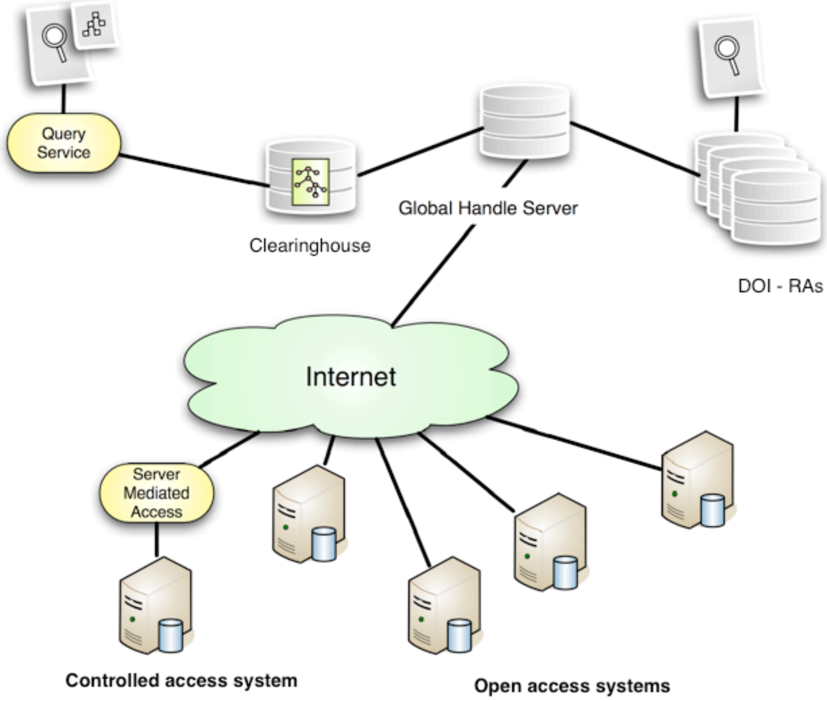

Figure 7. A proposed system using persistent identifiers and existing systems to permit tracking the use of genetic resources and monitoring compliance with the CBD-ABS regime. A DOI based service, acts as a clearinghouse in which PIDs are used to aggregate the entire documentation history of the use of a genetic resource for either commercial or non-commercial research. The system supports human and machine queries and facilitates the retrieval of all relevant documents from public and private databases, including the STM literature, patent and regulatory databases.

The selection of an appropriate PID for the CBD/ABS and related activities will be critical for its broad utility and community acceptance. However, it does not obviate the importance of carefully defining precisely what the identifiers point to, and what will be returned by queries of various types. It is possible to develop an identifier scheme that will provide a direct link to digital and paper copies of entire documents, such as PIC, MTAs, CoO and other relevant agreements. Likewise, if digital copies of ABS critical documents are marked up in a consistent format (*e.g.*, in XML using document specific DTDs or schemas), then DOI-aware applications can be readily applied to embed persistent links into the content that will permit tracking of genetic resources or parts of genetic resources in a future proof method, or do so on-the-fly. Means of tracking the transfer of materials and the corresponding agreements to third parties is also possible, in a manner that is consistent with the rights and obligations of all parties to the initial agreement or to subsequent agreements. Similarly, the ability to trace these genetic resources into the STM, general interest and patent literature is technically feasible. Reduction to practice will require a commitment of interested parties from different sectors (*e.g.*, government, industries, botanical gardens, museums, academia, etc) to define standards for the key documents that are instrumental to implementing the ABS. Business rules and policies also need to be established in concrete terms so that useful prototypes can be built and assumptions (technical, legal and social) tested and refined.

Creation of a trusted clearinghouse system for tracking and monitoring the use of genetic resources in a manner that complies with the provisions of the CBD is a challenging task, but not impossible. It is just one that will require that we draw upon a multiplicity of skills and knowledge to develop and implement a clearinghouse system that is enduring, efficient and trustworthy.

Useful Definitions

Identifier: strings of numbers, letters, and symbols that represent some *thing*.(Coyle, 2006). Identifiers are lexical tokens or names that denote things; a referent is the thing that is identified by an identifier.

Identifier system: functional deployment of identifiers in computer sensible form through assignment, resolution, referent description, administration, etc, which uses an Identifier in conjunction with some additional technical and/or social infrastructure in order to provide an application service for identifiers.

Interoperability: the ability of independent systems to exchange meaningful information and initiate actions from each other, in order to operate together for mutual benefit. In particular, it envisages the ability for loosely-coupled independent systems to be able to collaborate and communicate. At least three levels exist: Syntactic interoperability (the ability of systems to process a syntax string and recognise it as an identifier even if more than one such syntax occurs in the system); Semantic interoperability (the ability of systems to determine if two identifiers denote precisely the same referent; and if not, how the two referents are related); and Community interoperability (the ability of systems to collaborate and communicate using identifiers whilst respecting any rights and restrictions on usage of data associated with those identifiers in the systems).

Naming authority: authority that assigns names and guarantees their uniqueness and persistence. A naming resolution service corresponds to every naming authority and carries out the name resolution. A PI distributed system foresees that the responsibility of generation and resolution can be delegated to other institutions called sub-naming authorities who manage a portion of the name domain/space.

Namespace: an abstract container providing context for the items it holds and allows disambiguation of items having the same name (residing in different namespaces) (source: http://en.wikipedia.org/wiki/Namespace)

Object: any entity of interest in a transaction. A particular object identified by a specific identifier is the referent of that identifier. Since the object should persistently continue to be the same thing, it is necessary to explicitly state and define "what is" the referent of the identifier, defined by metadata and terminology from a common data dictionary to ensure that "what you mean is what I mean'' (see interoperability). Objects can be physical, digital, or abstract, *e.g.*, people, organizations, agreements, *etc*. (Bellini *et al.*, 2008)

Persistent identifier: an identifier for an object that uniquely identifies that object It can be used in services outside the direct control of the issuing assigner without a stated time limit. It will never be reassigned to any other object and will not change regardless of where the object is located or whatever protocol is used to access it. (Garrity, 2007)

Reference implementation: a software example of a specification which is intended to help others implement their own version of the specification or find problems during the creation of a specification.

Register: name association table between URNs and one or more URL.

Resolution: the process in which an identifier is the input (a request) to a network service to receive in return a specific output (object, metadata, *etc*.).

Resolution service: a service maintaining necessary infrastructure to provide *resolution* of a set of identifiers.

Unique identification: the specification by an identifier of one and only one referent (although one referent may have more than one identifier)

URI: A Uniform Resource Identifier is the generic set of all names/addresses that are short strings that refer to objects

URL: A Uniform Resource Locator is a URI that, in addition to identifying an object, provides means of acting upon or obtaining a representation of the object by describing its primary access mechanism or network "location".

URN: A Uniform Resource Name is a URI is a persistent, location-independent object identifier which is designed to make it easy to map other namespaces (that share the properties of URNs) into URN-space

Useful Resources:

Persistent identifiers

Digital Curation Centre (2005) *Proceedings of the DCC Workshop on Persistent Identifiers*, 30 June-1 July (Glasgow). URL: http://www.dcc.ac.uk/events/pi-2005/

ERPANET (2004) *Persistent Identifiers*, Final Report of the ERPANET Workshop on Persistent Identifiers, 17-18 June University College Cork, Ireland. Available at: PDF

Hilse, H.W.and Kothe,, J. (2006), *Implementing Persistent Identifiers* Consortium of European Research Libraries, November Available at: PDF

Dack, D. (2001) *Persistent Identification Systems* Report to The National Library of Australia http://www.nla.gov.au/initiatives/persistence/PIcontents.html

NISO (2006) *Report of the NISO Identifiers Roundtable*, 13-14 March 2006. Bethesda, MD: National Library of Medicine. Available at: PDF http://www.niso.org/news/events_workshops/ID-workshop-Report2006725.pdf

PILIN Team (2007) *Persistent Identifier Linking Infrastructure*, project final report Available at : PDF

URI

Berners-Lee, T., Universal Resource Identifiers in WWW, RFC 1630 (June 1994). http://www.ietf.org/rfc/rfc1630.txt

Berners-Lee, T., Fielding, R., and Masinter, L., Uniform Resource Identifier (URI): Generic Syntax, RFC 3986 (January 2005). http://www.ietf.org/rfc/rfc3986.tx t

Daniel, R. and Mealling, M., Resolution of Uniform Resource Identifiers using the Domain Name System, RFC 2168 (June 1997). http://www.ietf.org/rfc/rfc2168.txt

URL

Berners-Lee, T., Masinter, L., and McCahill, M., Uniform Resource Locators (URL), RFC 1738 (December 1994). http://www.faqs.org/rfcs/rfc1738.html

URN

Moats, R., URN Syntax, RFC 2141 (May 1997). http://www.ietf.org/rfc/rfc2141.txt

Internet Assigned Numbers Authority, 'URN Namespaces', Internet Assigned Numbers Authority website. http://www.iana.org/assignments/urn-namespaces

Sollins, K., and Masinter, L., Functional Requirements for Uniform Resource Names, RFC 1737 (December 1994). http://www.w3.org/Addressing/rfc1737.txt

Persistent identifiers: URN http://www.paradigm.ac.uk/workbook/metadata/pids-urn.html

URIs, URLs and URNs

Mealling, M., and Denenberg, R., *Report from the Joint W3C/IETF URI Planning Interest Group: Uniform Resource Identifiers (URIs), URLs, and Uniform Resource Names (URNs): Clarifications and Recommendations*, RFC 3305 (August 2002). http://www.ietf.org/rfc/rfc3305.txt

W3C, 'Naming and Addressing: URIs, URLs,', http://www.w3.org/Addressing/

PURL

Online Computer Library Center, PURL http://purl.org/ Includes links to overview documents and FAQs.

LSID

Atev, S. and Szekely, B. (2004) Build an LSID Resolution Service using the Java language. http://www.ibm.com/developerworks/opensource/library/os-lsid/

Swartz, A. (2002) RDF Primer Primer. http://notabug.com/2002/rdfprimer/

ARKS

California Digital Library, 'Archival Resource Key (ARK), *California Digital Library website*. http://www.cdlib.org/inside/diglib/ark/

Kunze, John A., *Towards Electronic Persistence Using ARK Identifiers* (July 1993). http://www.cdlib.org/inside/diglib/ark/arkcdl.pdf

Kunze, J., and Rodgers, R.P.C., *The ARK Persistent Identifier Scheme*, Internet Draft (23 August 2006). http://www.ietf.org/internet-drafts/draft-kunze-ark-14.txt

Handle System

Corporation for National Research Initiatives, 'Handle Resolver Service',.: http://hdl.handle.net/

Kahn, Robert, and Wilensky, Robert, A Framework for Distributed Digital Object Services (May 1995). http://www.cnri.reston.va.us/k-w.html

Sun, S., Lannom, L., and Boesch, B., Handle System Overview, RFC 3650 (November 2003). http://www.ietf.org/rfc/rfc3650.txt

Sun, S., Reilly, S., and Lannom, L., Handle System Namespace and Service Definition, RFC 3651 (November 2003). http://www.ietf.org/rfc/rfc3651.txt

Sun, S., Reilly, S., Lannom, L. and Petrone, J., Handle System Protocol (ver 2.1) Specification, RFC 3652 (November 2003). http://www.ietf.org/rfc/rfc3652.txt

DOI

The International DOI Foundation (IDF), *DOI System website*. http://www.doi.org/ including overviews, DOI handbook *etc*.

International DOI Foundation: *Key facts on DOI System* (2008) at http://www.doi.org/factsheets/DOIKeyFacts.html

The International DOI Foundation (IDF), 'Resolve a DOI', DOI System website. http://dx.doi.org

Paskin, N (2008) *DOI System*: article in third edition of the *Encyclopedia of Library and Information Sciences*, Taylor & Francis Group (in press). Preprint version available at http://www.doi.org/overview/080625DOI-ELIS-Paskin.pdf (Revised June 2008).

References

ABR (2009) Permits–accessing biological resources in Commonwealth areas. Available at http://www.environment.gov.au/biodiversity/science/access/permits/index.html

Ad Hoc open-ended working group on ABS. (2007) *2007 Final Report of the meeting of the group of technical experts on an internationally recognized certificate of origin* October 8, 2007. http://www.cbd.int/doc/?mtg=absgte-01

Bellini, E., Cirinnà, C., and Lunghi, M. (2008) *Persistent Identifiers for Cutural Heritage*. Fondazione Rinascimento Digitale: DPE Briefing Paper. Available at: PDF

Berners-Lee, T., Fielding, R. and Masinter, L. (2005) *Uniform Resource Identifier (URI): Generic Syntax* Internet Engineering Task Force (IETF) Request for Comments (RFC), RFC 3986. Available at: http://tools.ietf.org/html/rfc3986#page-7

Bourne PE, Fink JL (2008) I Am Not a Scientist, I Am a Number. *PLoS Computational Biology*, 4(12): e1000247 doi:10.1371/journal.pcbi.1000247

Broeder, D. (2007) *Persistent Identifiers*. Presentation at the DAM-LR Meeting, Lund University in Lund, Sweden December 18, 2007. Available at: PDF

CDL (2009) Name Assigning Authority / Name Mapping Authority Lookup Table at http://www.cdlib.org/inside/diglib/ark/natab

CENDI Persistent Identifiers Task Group. (2004) *Persistent Identification: A Key Component of an E-government Infrastructure*. Whitepaper. Available at: PDF

Clark, D., Sollin, K. *et al.*, (2003) *New Arch: Future Generation Internet Architecture*. Rome, NY: DoD. PDF

Clark, T. (2003) Identity and interoperability in bioinformatics. *Briefings in Bioinformatics,* 4(1): 4-6. doi:10.1093/bib/4.1.4

Clark, T., Martin, S. and Liefeld,T. (2004) Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics,* 5(1): 59-70. doi:10.1093/bib/5.1.59

Crossref (2008) Crossref Indicators on December 01, 2008. http://www.Crossref.org/01company/Crossref_indicators.html

Davidson, J. (2006) *Persistent Identifiers*. Briefing paper at University of Glasgow, October 18, 2006. http://www.dcc.ac.uk/resource/briefing-papers/persistent-identifiers/

Davidson, L., and Douglas, K. (1998) Digital Object Identifiers: Promise and Problems for Scholarly Publishing. *Journal of Electronic Publishing,* 4 (2). 10.3998/3336451.0004.203

Deutsche Nationalbibliothek (no date) *Persistent Identifier*. Retrieved November 15, 2008 from http://www.persistent-identifier.de/?link=204&lang=en

Garrity, G.M. (2007) *An Overview of Persistent Identifiers* presented at the IT Support for SMTA implementation, Rome Italy, February 14, 2007.

Garrity, G.M. (2006) *Digital Object Identifiers as a technology Implementation of a full working prototype The NameforLife Model*. Presented at GUID-1 workshop, Durham, NC. February 1-3, 2006. Available at: PPT

Garrity, G.M. and Lyons, C. (2003) Future-proofing biological nomenclature. *OMICS: A Journal of Integrative Biology*, 7(1): 31-33. doi:10.1089/153623103322006562

Hagedorn, G. (2006) *Why We Should Not Use LSIDs* Retrieved November 17, 2008, from the TDWG Wiki http://wiki.tdwg.org/twiki/bin/view/GUID/WhyWeShouldNotUseLSIDs

Handle (2009) www.handle.net

Handle System Client Extension (2009) http://www.handle.net/hs-tools/extensions/mozilla_hdlclient.html

Hilse, H., and Kothe, J. (2006) *Implementing Persistent Identifiers: overview of concepts, guidelines and recommendations*. European Commission on Preservation and Access (ECPA) report 18. Available at: PDF

IDF (2006) *DOI Handbook*. doi:10.1000/182

IDF (2008) *Key facts on DOI System*. Available at http://www.doi.org/factsheets/DOIKeyFacts.html

Kahn, R., and Wilensky, R. (1995) *A Framework for Distributed Digital Object Services,* Technical Report tn95-01. Corporation for National Research Initiatives. Available at: http://www.cnri.reston.va.us/k-w.html. (re-published, with an additional introduction by the authors in International Journal on Digital Libraries (2006) 6(2): 115-123. doi: 10.1007/s00799-005-0128-x

Kunze, J. (2001) *The ARK Persistent Identifier Scheme* February 22, 2001. Internet Engineering Task Force (IETF) Internet-drafts. Available at: http://tools.ietf.org/html/draft-kunze-ark-00

Kunze, J. (2003) *Towards Electronic Persistence Using ARK Identifiers*. Retrieved November 18, 2008 from. http://www.sspnet.org/documents/149_Kunze.htm

Kunze, J. (2006) *Noid* (*Nice Opaque Identifiers) Minting and Binding tool: overview and technical specification.* Available at: PDF

Kunze, J. (2008) Inside CDL: ARK (Archival Resource Key). Retrieved November 18, 2008 from http://www.cdlib.org/inside/diglib/ark/#ark

Kunze, J. (2008) *Persistent Identifier Principles and Practice* presented at the International Conference on Dublin Core and Metadata Applications, in Berlin, Germany September 24, 2008. Available at: PDF

Kunze, J., and Rodgers, R. (2008) *The ARK Identifier Scheme.* Internet Engineering Task Force (IETF) Internet-drafts. Available at: http://tools.ietf.org/html/draft-kunze-ark-15

Kunze, J., and Rodgers, R. (2001) *The ARK Persistent Identifier Scheme*. Internet Engineering Task Force (IETF) Internet-Drafts. Available at http://tools.ietf.org/html/draft-kunze-ark-01

Lannom, L. (2000) *Handle System Overview*. 66th IFLA Council and General Conference, 23-28 August. Jerusalem, Israel. Available at: http://www.ifla.org/IV/ifla66/papers/032-82e.htm

Lannom, L. (2008) *Introduction and Handle Update*. Handle System Workshop, 17 June, 2008. Brussels, Belgium. Available at: PPT

Marieke , G. (2005) *An Introduction To Persistent Identifiers* QA Focus briefing document no. 80. UKOLN. Available at: http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-80/html

Moats, R. (1997) *URN Syntax.* Internet Engineering Task Force (IETF) Request for Comments RFC 2141. Available at http://www.ietf.org/rfc/rfc2141

NISO (2006) *Report of the NISO Identifiers Roundtable*. NISO Identifier Roundtable March 13-14, 2006 National Library of Medicine, Bethesda, MD Available at: PDF

NISO (2004) *Understanding Metada*. Bethesda: NISO Press. Available at: PDF

Object Management Group (2003) *Life Sciences Identifiers, OMG Adopted Specification* Available at: PDF

Paradigm (2008) Administrative and preservation metadata. Retrieved November 05, 2008 from http://www.paradigm.ac.uk/workbook/metadata/index.html

Paskin, N. (1997) Information Identifiers. *Learned Publishing* 10(2): 135-156.

Paskin, N. (1999) Toward unique identifiers. *Proceedings of the IEEE* 87(7): 1208-1227. doi: 10.1109/5.771073

Paskin, N. (2003) Components of DRM Systems: Identification and Metadata. In E. Becker *et al. Digital Rights Management*: *Technological, Economic, Legal and Political Aspects in the European Union* in the series Lecture Notes in Computer Science. p. 26-61 New York: Springer-Verlag. Available at: PDF

Paskin, N. (2004) *The development of persistent identifiers* presented at the ERPANET Persistent Identifiers seminar, Cork, Ireland June 17-18, 2004. Available at: PDF

Paskin, N. (2005a) Digital Object Identifiers for scientific data. *Data Science Journal*, 4: 12-20. doi:10.2481/dsj.4.12

Paskin, N. (2005b) *DOI* presented at the DCC Workshop on Persistent Identifiers, University of Glasgow, July 1, 2005. http://www.dcc.ac.uk/events/pi-2005/

Paskin, Norman (2006a). "Identifier Interoperability: A Report on Two Recent ISO Activities", *D-Lib Magazine*, 12 (4). April 2006. doi:10.1045/april2006-paskin

Paskin, N. (2006b) Naming and meaning of Digital Objects. *Proceedings of 2nd International Conference on Automated Production of Cross Media Content for Multi-channel Distribution* (p. 42-49). Leeds: University of Leeds. Available at: PDF

Paskin, N (2008a) DOI System. In M. Bates & M. N. Masck (Eds), *Encyclopedia of Library and Information Sciences (*3r ed.*).* Taylor & Francis Group (in press). Preprint version available at PDF (Revised June 2008)

Paskin, N. (2008b) *Identifier Interoperability*. Fondazione Rinascimento Digitale: DPE Briefing Paper. Available at: PDF

PILIN project (2007) *Using URLs as Persistent Identifiers.* Retrieved January 2009, from http://resolver.net.au/hdl/102.100.272/DMGVQKNQH

Pogrebin, R. (1998) For $19.95, Slate Sees Who Its Friends Are. *New York Times*, March 30, 1998.

Rust, G. and Bide, M. (2000) *The <indecs> Metadata Framework: Principles, model and data dictionary*. No. WP1a-006-2.0. Indecs Framework Ltd. Available at: PDF

Salamone, S. (2002) LSID: An Informatics Lifesaver. *Bio-itworld.com*, January 12, 2002. Available at: http://www.bio-itworld.com/archive/011204/lsid.html

Sowa, J. F. (2000) *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Pacific Grove, CA: Brooks Cole Publishing Co.

STD-DOI (2008) Publication and citation of Scientific Primary Data at http://www.std-doi.de/front_content.php

Sun, S., Lannom, L., and Boesch, B. (2003) *Handle System Overview.* Internet Engineering Task Force (IETF) Request for Comments (RFC), RFC 3650. Available at: http://www.ietf.org/rfc/rfc3650.txt

Sun, S.; Reilly, S., and Lannom, L. (2003) *Handle System Namespace and Service Definition*. Internet Engineering Task Force (IETF) Request for Comments (RFC), RFC 3652. http://hdl.handle.net/4263537/4068

Szekely, B. (2003) Build a life sciences collaboration network with LSID. Retrieved November 2008 from http://www.ibm.com/developerworks/opensource/library/os-lsid2/?S_TACT=105AGY82&S_CMP=GENSITE

Szekely, B. (2006) *LSID as a Technology: Overview, participation and related projects*. Presented at TDWG GUID Workshop at Durham, NC February 1-3, 2006 Available at: PPT

Tonkin, E. (2008) *Persistent Identifiers: Considering the Options*. *Ariadne* , 56 (July 10, 2008) Available at http://www.ariadne.ac.uk/issue56/tonkin/

Van Brabant, B., Gray, T., Verslyppe, B., Kyrpides, N., Dietrich, K., et al., (2008) Laying the Foundation for a Genomic Rosetta Stone: Creating Information Hubs through the Use of Consensus Identifiers. *OMICS: A Journal of Integrative Biology*, 12(2): 123-127. doi:10.1089/omi.2008.0020

Veretnik S, Fink JL, Bourne PE (2008) Computational Biology Resources Lack Persistence and Usability. *PLoS Computational Biology* 4(7): e1000136 doi:10.1371/journal.pcbi.1000136

Ward, N., and Macnamara, D. (2007) *Workshop 7: Persistent identifiers: Being persistent and pervasive in the national interest*. Presentation at the eResearch Australasis on June 29, 2007 in Sydney, Australia. http://www.eresearch.edu.au/ws7

CONCLUSION AND RECOMMENDATIONS

We appreciate the opportunity that the Secretariat has provided us to review the agreed-upon subject matter in-depth. For us, it is clear that advancements in the field of genomics and related disciplines (*e.g.,* proteomics, transcriptomics, systems biology, bioinformatics) of the past decade have had far reaching effects and have permanently altered the way in which biology is and will be practiced in the future. While these advancements have been driven largely by needs in medicine and public health, the technologies are adaptable to biodiversity research. The key driver has been the rapid developments in sequencing technology coupled with the precipitous decline in the cost of producing sequence data. The possibility of having a detailed genomic signature for any organism of interest at low cost opens many new opportunities and challenges.

We have also gained an increased awareness of the tools and techniques that permit expression of foreign genes in heterologous expression systems as well as the possibilities and consequences that these approaches bring to discussions about sustainable use of genetic resources. Whereas whole organisms (*e.g.*, plants, animals, bacteria) were once the subject of interest as a source of products, processes, and leads for new chemical entities, now it is just as likely that genome sequence data is the subject of greatest interest. Contemporary discovery methods rely on genome mining and gene probes to isolate the genes and pathways of interest. Once in hand, those genes can be easily modified and expressed in well understood hosts, including semi-synthetic chimeras today and engineered synthetic life forms that soon will be custom designed to carry out metabolic reactions that do not occur in nature. The genome data is now as valuable as, or even more valuable than, the organism, and vast amounts of sequence data are readily available from publicly available databases.

Through this exercise, we have also become increasingly aware of the marked difference in the pace of change between science and technology, and social policy and law. There are other examples in recent history, from which we can draw parallels (*e.g.*, information technology, telephony) that may inform us as to likely outcomes. New technologies (especially disruptive ones) tend to have immediate effects as they pervade targeted markets and challenge underlying assumptions on which rules, regulations, policies and laws are based. Changes in the latter are necessarily slower as the consequences of change are not always evident early on. However, the technological changes are invariably permanent and irreversible. We are also aware that successful implementation of any computing system is dependent on more than technical issues. Human and social issues need to be factored into the design to ensure usability, acceptability and ultimately trust in any system that may ultimately be deployed to facilitate the ABS regime. This will require supporting data policies and business rules that meet the needs and protect the interests of all parties involved.

The Secretariat and the COP are to be commended for their pragmatism and understanding that the changes we have discussed in our report must be considered when devising an ABS regime that will be workable in years to come. To that end, we offer the following recommendations based upon our findings.

Recommendations

1. Considerable discussion has occurred concerning the documents that are deemed essential in establishing the rights and obligations of providers and users of genetic resources. PICs, MTAs, MATs and CoOs establish precisely what genetic resources are being provided, the parties involved, the terms of use, the intended types of activities permissible, and the rights and obligations of each party. Despite their central importance to the process, such documents do not yet appear to exist in any standardized manner.

   The Parties urgently need to resolve this matter. These documents serve as the triggering event for tracking genetic resources and must be digitally bound to each and every resource (in the form of links made through PIDs). We recognize that these documents contain confidential information that is specific to a particular agreement, but these documents also contain information that is common to all like documents. This common information comprises a minimal set of descriptors that are important for tracking genetic resources and may, in time, be required for legal and regulatory purposes. We recommend that the parties consider a central registry for recording this minimal information, such as a digital certificate of compliance, which could be used for a variety of legal and regulatory purposes to indicate that these agreements are in place without violating the confidentiality of the agreement.

2. There is a general tendency within some communities to eschew well-established tools and solutions that are available "off-the-shelf". The most common argument used to justify this rationale is expense. However, the costs of developing and maintaining computing systems with a large user-base, a requirement for near constant up-time, redundancy, and ongoing curation are rarely factored into such justifications. Nor is the cost of potential failure, should funding no longer be available or key individuals depart from a project. The anticipated tracking system for genetic resources must be developed from the outset as a commercial-grade resource that will meet the needs of all parties that will come to rely on its continued functioning, whether they are in the public or private sectors. As noted above, the anticipated system may also play an important role in policies, rules, regulations and laws at the national and international level.

   We encourage the parties to carefully examine existing identifier systems and select one that is widely used, interoperable, and well supported by a large and diverse user community. It is significantly less expensive to modify an existing identifier system than to build one. A conservative estimate of the investment made to create, deploy, support and extend the Handle server, including the DOI system and the network of registration agencies and application developers is in the range of $15-20 million US. The costs associated with other systems are perhaps lower, but none of the other identifier systems are in widespread use.

   We encourage the parties to carefully consider the concept of persistence. As noted above, persistence is not a technical problem. Rather it is a social and business problem. To ensure persistence and reliability, policies will be needed to ensure proper use of the system and mechanisms for sustained funding. It is anticipated that the genetic resource tracking system will have various classes of users, each with different rights and obligations. These policies need to be developed to ensure that data integrity is not compromised. Business models also need to be considered to defray the cost of operation into the future.

3. The concept of sustainable use of biodiversity is broad and vague in that it does not adequately describe what constitutes a unit of biodiversity. In certain contexts (*e.g.*, biological inventories, ecological studies, taxonomic research), the simple assignment of a species or higher taxon name may be adequate to meet objectives, as one specimen is essentially identical to the next. However, there are many cases in which the desired properties may be found in only a small number of individuals, or the gene or pathway of interest may not be restricted to a single taxon. In such cases, agreements between providers and users may encompass many hundreds or thousands of unidentified organisms or materials for metagenomic analysis (*e.g.*, pharmaceutical or enzyme screening). Under such circumstances, it is a specific individual that must be tracked and taxonomic information may be of little or no predictive value.

   We recommend that the parties carefully consider the needs of all providers and users and recognize that the proposed tracking system is not merely a research tool to fulfill the needs of ecologists or taxonomists. Rather, the current and future needs of a much broader community must be considered and the granularity of the information is that tracked is likely to be variable. The correct associations between resources and associated documents, reports in the STM and gray literature (including regulatory, patent, policy and other non-indexed content relevant to CBD ABS objectives), and data from various sources must be established and maintained to ensure that the rights and obligations of providers and users are fulfilled.

4. The most common way in which data and information are disseminated today is through Web portals. Content and web services can be readily delivered to anyone with an Internet connection and a browser, whether it is a computer, a handheld device or mobile phone. Whereas accessibility to a reliable Internet connection was once a problem in some parts of the world, that problem is rapidly diminishing in significance. Usability of applications and reliability of data (including data provenance) are, however; becoming more problematic, particularly in the life sciences.

   We encourage the Parties to consider the potential hazards of data abuse by common practices such as screen scraping, redirection, and wholesale data harvesting. Such practices can diminish the trust in the resource as the underlying data can become compromised. This problem can be mitigated through use of persistent identifiers, most notably DOIs. The underlying business policies and digital rights management that are part of the DOI system, along with concept of identifier ownership and responsibility, provide an incentive and mechanism to address this problem. Lightweight applications that use well managed persistent identifiers can be readily deployed using browser-based applications that can be used to verify user rights and control access to data and other web services. Applications can also be built that permit trusted partners or subscribers to transclude data or other received data feeds on a push or pull basis.

5. Systems development is an expensive and time consuming undertaking and requires careful consideration of data types and data structures, end user needs, interfaces and the types of services to be provided. The process is typically iterative, and initial models and concepts often prove lacking in some way. Prototype applications provide an expedient way to test concepts and assumptions and to gain valuable insight into workable solutions that can be deployed on a larger scale.

   We recommend that one or more prototypes be developed to validate underlying concepts and to provide guidance to the Parties in further defining critical elements that need to be accommodated in a fully operational system. We further recommend that such a system be developed in conjunction with a national competent authority that has already developed PICs, MTAs, MATs and CoOs so as to provide meaningful test cases.