

**Note: The summary is only author's observation from the submitted presentation/paper and may not reflect to the presenter's intent.
Draft as August 10, 2010**

US DPIF Workshop: Content Organization

Submission Questions	Federated Search and Good Databases (DOE)	NASA's Earth Science Data Systems - Lessons Learned and Future Directions (NASA)	NARA Electronic Records Archives (ERA) Lessons Learned and Future Direction (NARA)	Creating a Digital Repository (LoC)	Archiving Strategy for EROS Center (USGS)	Digital Preservation Framework, Goals, and Challenges at GPO (GPO)	Safekeeping, Finding, and Sharing SI's Digital Diamonds (Smithsonian)	AWIPS Approach to Exponential Increase in NOAA Data Volume (NOAA)	Improving and Strengthening Inter-Institutional Preservation (SDSC)	Chronicles in Distributed Digital Preservation (MetaArchive)
Background Institution/ Department	To advance science and sustain technological creativity by making R&D findings available and useful to Department of Energy (DOE) researchers and the public	Study Earth from space to advance scientific understanding and meet societal needs via core and community data system elements : <ul style="list-style-type: none"> • Core – manage Earth science satellite mission data, airborne instrument data and field campaign/ in situ measurement data robustly to ensure continuity of research, 	ERA is the National Archives and Records Administration (NARA)'s strategic initiative to preserve and provide long-term access to uniquely valuable electronic records of the U.S. Government, and to transition government-wide management of the lifecycle of all records into the realm of e-government.	<ul style="list-style-type: none"> • Copyright regulator for the United States • Serves the research needs of the Members of Congress and their staffs - Congressional Research Service - Law Library of Congress 	The U.S. Geological Survey's Earth Resources Observation and Science Center has the responsibility to acquire, manage, and preserve our Nation's land observations. These records are obtained primarily from airplanes and satellites dating back to the 1930s.	The core mission of Keeping America Informed, dated to 1813 when Congress determined to make information regarding the work of the three branches of Government available to all Americans. The U.S. Government Printing Office (GPO) provides publishing & dissemination services for the official & authentic government publications to Congress, Federal agencies, Federal	<ul style="list-style-type: none"> • World's largest museum & research complex • 19 museums and the National Zoo • Over 30 million visitors in 2009 • 188 million Website visitors • 9 science centers including observatories, conservation and tropical research • 136 million+ collection objects, artworks and specimens 	To understand changes in weather, climate, oceans, and coasts, to share that understanding with others, and to use it to manage natural resources – all to meet our nation's economic, social, and environmental needs	The San Diego Supercomputer Center and the UCSD Libraries with their partners NCAR and UMIACS have created Chronopolis to address these issues. Chronopolis is a national center for the management, long-term preservation, and promulgation of national digital assets.	MetaArchive Cooperative has built a trustworthy digital repository to provide for the long-term care of digital materials. It is a community-owned, community-led initiative. Its collaborative networks are comprised of libraries, archives, and other cultural memory organizations that seek to cooperatively preserve their digital materials, not by outsourcing to other organizations, but by actively

		<p>access, and usability</p> <ul style="list-style-type: none"> • Community – support data system evolution with technological innovations and develop data products to complement Core capabilities 				<p>depository libraries, & the American public.</p>				<p>participating in the preservation of their own content.</p>
Offering	<p>Federated search allows users to search multiple data sources simultaneously, in parallel, using a single query from a single user interface. OSTI offers federated search to patrons as a free aggregator of multiple government R&D-related databases.</p>	<p>Provide end-to-end capabilities to deliver Earth science data and information products to users</p> <p>Provide “one-stop-shopping” search and access to datasets in distributed Data Centers based on specific Earth science disciplines, as well as tailored access to datasets individual Data Centers.</p>	<p>NARA has preserved and provided access to the records of the United States of America. Records help us claim our rights and entitlements, hold our elected officials accountable for their actions, and document our history as a nation. In short, NARA ensures continuing access to the essential documentation of the rights of American citizens and the actions of their Government.</p>	<p>See Size under Content</p>	<p>USGS/EROS archives remotely sensed images of the Earth's land surface. These data are acquired by civilian satellites and aircraft and used to study a wide range of natural hazards, global environmental change, and economic development and conservation issues. EROS staff members manage and distribute these data to scientists, policy makers, and educators worldwide.</p>	<ul style="list-style-type: none"> • Ensure permanent, no fee public access to electronic publication • Ensure the public has access to Federal government information in all formats • Expand public access to Federal information that can be easily searched, retrieved, downloaded and indexed 		<p>Provide climate, water, weather forecasts and warnings to protect life and property and enhance the economy</p>	<p>Chronopolis provides comprehensive model for the cyberinfrastructure of collection management, in which preserved intellectual capital is easily accessible, and research results, education material, and new knowledge can be incorporated smoothly over the long term. Chronopolis provides trusted preservation environments for academic institutions and research projects, with the goal of</p>	<p>MetaArchive Cooperative provides low-cost, high-impact preservation services to help ensure the long-term accessibility of the digital assets of universities, libraries, museums, and other cultural memory organizations.</p>

Audience									long-term collection management, preservation, and knowledge generation. About	
	<ul style="list-style-type: none"> • General public • Scientific community • Etc. 	Scientific and applications users, general public, other federal agencies, educators, international partners, value-added service providers. <ul style="list-style-type: none"> • 	<ul style="list-style-type: none"> • General public and researcher 	<ul style="list-style-type: none"> • General public and researcher 	General public and researcher	General public and other Federal agencies			<ul style="list-style-type: none"> • Institutional repositories • university libraries • other cultural heritage institutions 	Universities, libraries, museums, and other cultural memory organizations
Impact	Benefits	<ul style="list-style-type: none"> • Speed-up technology development • Enable knowledge discovery • Etc. 	<ul style="list-style-type: none"> • Enabling climate modeling, global change analyses, disaster response and impact analyses • Supporting near real-time applications in many societal benefit areas (agricultural efficiency, air quality, aviation safety, carbon management, coastal management) 	<ul style="list-style-type: none"> • Be able to locate any legal or historical records of the US 	<ul style="list-style-type: none"> • Safeguard culture heritage 	The ability to compare landscapes from the past with current information enables change analysis at local and global scales.			<ul style="list-style-type: none"> • Geographic replication for preservation • Technology planning • Advanced file management 	<ul style="list-style-type: none"> • Reduce short-and-long-term costs • Decentralize activities • Decrease dependence on 3rd party solutions

		<p>t, ecosystems, disaster preparedness, energy forecasting, homeland security, invasive species, public health, and water management.)</p> <ul style="list-style-type: none"> • Maintaining long-term, consistent Earth science data records relevant to climate and other terrestrial phenomena 								
Consequence if content not preserve	<ul style="list-style-type: none"> • Loss of valuable info and knowledge • Duplicating effort 	<ul style="list-style-type: none"> • Loss of data would have a negative impact on future verifiability of conclusions from global change analyses • Data from NASA missions are records of physical phenomena measured at specific times and, if lost, can 	<ul style="list-style-type: none"> • Loss of US historical records 	<ul style="list-style-type: none"> • Destroy culture history 	<ul style="list-style-type: none"> • Loss of observations, land cover, and land use historical data 				<ul style="list-style-type: none"> • Loss of valuable research data • Waste of research dollars 	Loss of our digital cultural heritage

<p>Benefits of using stds DP</p>	<ul style="list-style-type: none"> Utilize common DP methodologies Improve access via standard interfaces 	<p>never be recreated.</p> <ul style="list-style-type: none"> Enhance operation and management of data Improve communication / interface between internal and external systems 	<ul style="list-style-type: none"> Provide long term access to electronic records Ensure system evolvability by using flexible and standard system workflow and business rule engine. 	<ul style="list-style-type: none"> Enhance operation and management of data Improve better communication / interface between internal and external systems 	<ul style="list-style-type: none"> Provide consistent appraisal process Enhance operation and management of data Improve data steward accountability 				<ul style="list-style-type: none"> Ability to share data among systems, increasing preservation services for data providers 	<p>Low cost, high-impact preservation services without dependency on 3rd party or vendor solutions</p>
<p>Content Size</p>	<ul style="list-style-type: none"> More than 230,000 full-text documents Over 5 million scientific e-prints More than 2 million publicly available citations More than 24,000 patents Approximately 20,000 DOE current project summaries More than 500 websites & databases Conference papers & proceedings And more 	<ul style="list-style-type: none"> 4.2PB total archive volume 1.8TB daily growth >4000 unique data products >910K distinct users of EOSDIS data and services > 1M web site visits of 1 minutes or more >254M end user distribution products 6.7TB/day of end user average daily distribution volume (All of the above are metrics for FY 	<ul style="list-style-type: none"> 40M email messages for Clinton Admin. 25M electronic diplomatic messages 54M images from electronic official military personnel files annually 600M – 800M image files (2000 census) 	<p>World's largest archive:</p> <ul style="list-style-type: none"> 142M items 32M books in 470 languages Largest rare book collection in North America World largest collection of maps and sheet music 62M manuscripts 6.25M non-print media items (audio, video, film, electronic gaming and instruction). Adding 100K – 200K items per year Tape type: 	<p>Over 40,000,000 individual images make up EROS' digital collections occupying 1.6 petabytes of storage. An additional 8 million frames of analog photography exists on 100,000 rolls of film. EROS maintains at least two copies for all long-term digital records with the total electronic archive managed being over 4 petabytes.</p>			<p>l</p> <ul style="list-style-type: none"> Inter-university Consortium of Political and Social Research – preservation copy of collections including 40 years of social science data and Census California Digital Library – political and government web crawls, Web-at-risk collection SIO Explorer – data from 50 years of research voyages NCSU Libraries -- state and 	<p>Network: 254 TB and growing with 17 geographically distinct nodes in 13 states/districts and 3 countries</p>	

		2009)		Storage Tek robot with 9800 slots using T10000B (T1B) tape					local geospatial data	
Scope	<ul style="list-style-type: none"> • Full text • Scientific ePrint • Citations • Patents • DOE project summaries • Websites & databases • Conference papers & proceedings 	<p>Data derived from imagers, sounders, active and passive sensors on board satellites; airborne instruments and in situ measurements</p> <p>Data from US and international instruments on board US and international partners' satellites</p> <p>Raw data and derived digital scientific "standard" and near-real-time data products covering many Earth science disciplines – atmospheric, oceanographic, cryospheric, seismic, land processes and solar radiance</p>		See above under Size	The majority of EROS' holdings are images obtained from satellites or airplanes. Both of these sources, satellite or airplane, have analog or digital ties.				<ul style="list-style-type: none"> • Content-agnostic system, designed to house data from a variety of sources 	<ul style="list-style-type: none"> • Archives based on subjects and genres • Current archives include: <ul style="list-style-type: none"> - Southern Digital Culture Archive - ETD Archive (with NDLTD) - Early and Modern Literature Archive - Newspaper Archive - TransAtlantic Archive - General Archive

Distribution mechanism	Web interfaces	<ul style="list-style-type: none"> On-line protocols (ftp, http, etc.) and services 	<ul style="list-style-type: none"> On-line Physical media 	<ul style="list-style-type: none"> On-line?? Physical media?? 	Web interfaces				<ul style="list-style-type: none"> Online via a number of standard tools 	Web-based
Technologies (metadata, file format, packaging, etc.)	Format: <ul style="list-style-type: none"> Dublin Core MARC XML OAI 	Metadata: <ul style="list-style-type: none"> ECS (EOSDIS Core System) metadata FGDC-STD-012-2002 metadata for remote sensing data EOS Clearing House (ECHO) metadata ISO 19115 for geographic information and services metadata File format: <ul style="list-style-type: none"> Primary HDF-EOS HDF4 and HDF5 NetCDF GeoTIFF CEOS Binary Others... 	<ul style="list-style-type: none"> XML PREMIS data model 	Digitize everything <ul style="list-style-type: none"> Files should be able to stand alone without external reference (lots of metadata wrapped in the files) Digitize items in their original quality <ul style="list-style-type: none"> Audio – at highest bandwidth in the recording Video – native resolution Film – scanned at resolution equal to the highest available on film 	FGDC Content Standard for Geospatial Data. Tiff or GeoTiff for most distribution formats.		The Core Image Model: 79 metadata elements grouped into 5 main categories: <ul style="list-style-type: none"> Administrative elements (12) Descriptive elements (17) Technical elements (11) Rights Management elements (9) Embedded elements (29): IPTC or XMP 		<ul style="list-style-type: none"> Use the Storage Resource Broker (SRB) for data management (moving to iRODS) Use BagIt file packaging format and SRB tools to ingest and transfer data Use Auditing Control Environment (ACE) for integrity checking 	<ul style="list-style-type: none"> Ingest SIPs from the web or via OAI-PMH All open source tools, including our Conspectus database (records metadata about all AIPs) Working with the BagIt spec (LC/CDL) for AIP exchanges with other digital preservation service groups Tools <ul style="list-style-type: none"> enable content ingest from all leading repository systems (including DSpace, CONTENTdm, DigiTool, ETD-db, Fedora, etc.) Recovery pathways

<p>System Architecture</p>	<p>Using Connectors to Communicate with Databases</p>	<ul style="list-style-type: none"> • Geographically distributed (across the U.S.) system of systems • Scalable commodity hardware • Combination of custom and off-the-shelf software • Well-defined inter-system interfaces and API's 	<ul style="list-style-type: none"> • System utilizing commodity hardware • Software architecture based on SOA paradigm for system evolvability 	<p>200 TiB SAN</p> <ul style="list-style-type: none"> • Staging area for transmission to backup site and the tape library • Backup site has identical SAN & tape library <p>Tape library</p> <ul style="list-style-type: none"> • StorageTek robot with 9000 slots currently installed; 37,500 planned by 2013 • Currently using T10000 tapes with 1TiB/tape current capacity (9PiB available; 37.5PiB without refresh) • Upgrade path to ~48TiB/tape by 2019 	<p>Wide spectrum of hardware (instruments) and software components</p>		<ul style="list-style-type: none"> • Artertia DAMS 6.8 SP4: Application Server Sun T5220, 8 Core T2 Processor running Solaris 10 • DAMS Database Server (Asset Metadata): Server Sun E2900 with Oracle 10 Enterprise • Asset Repository: EMC Clarion SAN (Currently 22 TB) • Hot Folder Mount: NetApps FAS3140 • Daily backup to Tivoli Archive 		<ul style="list-style-type: none"> • Multiple storage systems around the United States of (intentionally) different vendors and configurations 	<p>LOCKSS-based network with data management tools layered on by the MetaArchive Cooperative</p>
<p>OAIS relevant</p>	<p>??</p>	<p>Yes. Parts of the system follow OAIS reference model. Examining other parts for compliance and/or functional mapping</p>	<p>YES: the entire reference model of OAIS</p>	<p>??</p>	<p>Being considered.</p>				<p>Yes. OAIS compliant system.</p>	<p>Conforms to the OAIS model. Conforms to all 84 criteria of TRAC (self-audit completed 2010)</p>

Access	Methods	Web-based federated search requires only that the pertinent data be published in a web accessible database	<ul style="list-style-type: none"> • Most data are available via ftp • Some data in near-line robotic tape • on-line services for searching, subsetting, reprojection, mosaicing, format conversion, statistical analyses, visualization etc. 	<ul style="list-style-type: none"> • Provide search and discovery of electronic records based on commercial search engines. • Provide search of archival business objects based on XML database. 	<ul style="list-style-type: none"> • Try to make data on-line as much as possible 	All data are open and free to the user through web-based interfaces.				Access via a web-based data portal.	Designated Community may have access to content whenever needed via WARC-based mechanisms
	Restrictions	None	All data are available at no cost to all users except where agreed upon with international partners	<ul style="list-style-type: none"> • Open public access system opened to general public. • Business archival application used by federal agencies and NARA staff. 	access is governed by Copyright law:	Vast majority are public domain.				Data only available to data providers. NO GENERAL ACCESS.	Because content is largely restricted by IP concerns, the Designated Community is the Producer
Management	What content to preserve	<ul style="list-style-type: none"> • Scientific information (see Scope under Content) 	<ul style="list-style-type: none"> • Observation data from mission instrument (raw data, derived key data products, geo radiance data, etc.) • Validation field campaign datasets and inter- 	<ul style="list-style-type: none"> • Anything to do with the US legal and historical records 	Everything	Formal appraisal process in place to ensure existing and collections offered to or sought by USGS meet our agency mission.				Any content desired by the data providers.	Anything, but current focus is mainly on digital cultural heritage

		<p>comparisons with other instruments</p> <ul style="list-style-type: none"> • Other agencies' ancillary datasets • Results from derived high-level products, applications and research • Pre-flight or preoperational performance measurements • Instrument / sensor calibration data and method, etc. 								
Duration	Indefinitely	Irrecoverable data need to be preserved indefinitely. Processes need to exist for determining duration of preservation for other data, derived products, etc.	Forever	Data must last at least 4000 years (cf. "Life of the Republic...")	All records have set disposition dates on an approved NARA scientific records disposition schedule. A subset of these records are part of the National Satellite Land Remote Sensing Data Archive which by law will be retained by and with the USGS forever.				Indefinitely	Indefinitely

<p>Lifecycle strategies</p>	<p>??</p>	<ul style="list-style-type: none"> Periodically refresh media including 'touching' all data Budget for hardware refresh every three years Metadata is a key cost driver <ul style="list-style-type: none"> - needs to be continually reconciled and updated - changes with each new data model - websites are useless without good metadata Science discipline expertise is required for management of data 	<p>Transformation of electronic records. Media refresh.</p>	<p>Migrations will happen depending on technology refreshes. Currently anticipated to be every 5-10 years.</p>	<p>Utilize the USGS EROS Appraisal process and the NARA-USGS scientific records disposition schedule to determine the long-term value and thus the lifecycle of our records.</p>	<ul style="list-style-type: none"> Refreshment Migration Emulation 			<ul style="list-style-type: none"> Periodic technology refreshes; periodic content refreshes. 	<p>Data refreshing</p> <p>Migration</p> <p>Technology infrastructure refreshed every three years</p> <p>Work closely with Producers to assess and refine their data management practices to ensure that SIPs are as robust as possible</p>
<p>Standards Used</p>	<p>Open Database Connectivity (ODBC)??</p>	<p>OPeNDAP, OGC standards where they apply Also see Technologies under Content</p>	<p>XML XFORM Web Services BPEL XQuery, XPath</p>	<p>Use industry standards as much as possible</p>	<p>MXF for Moving Image content; BWF (RF64) for audio content</p>			<p>Keys to success - common standards and common reference implementation:</p> <ul style="list-style-type: none"> Machine to machine exchange essential to operations Meteorologi 	<p>Industry standards wherever possible.</p>	<p>Industry standards, including OAIS</p>

								<p>cal and oceanographic data are multidimensional, continually evolving, highly spatial and highly temporal in nature</p> <p>Open Geospatial Consortium (OGC)</p> <p>Meteorology</p> <ul style="list-style-type: none"> • Developing OGC standards and profiles for effective interoperability for web services and content across wider geospatial domain • Developing oceanographic, meteorological and climatological data, metadata, and web services interoperability standards 		
Protection Sensitivity (Privacy Confidentiality Security Intellectual Property)	<ul style="list-style-type: none"> • Depends on database content • Depends on respective content 	<ul style="list-style-type: none"> • None for scientific data and derived products with the 	<ul style="list-style-type: none"> • Yes, depends on the nature of the records 	<ul style="list-style-type: none"> • No Copyrighted content can be released outside of Library 	<ul style="list-style-type: none"> • Vast majority of our science records are public domain. We 				<ul style="list-style-type: none"> • No general access to data in system • System does no accept 	<p>Highly secure network with no access provided to anyone except the designated</p>

	owner's policy	exception of those governed by international agreements <ul style="list-style-type: none"> User-related information is protected in compliance with privacy regulations Some documentation is protected as SBU or ITAR-sensitive 		systems	do manage some copyright data and control privacy information.				sensitive data	community (which is usually the same as the Producer)
Challenges	Issues	<ul style="list-style-type: none"> It can't be difficult to connect non-well defined databases Scientific disciplines have different ways of looking at the data and different vocabularies . Need flexibility and tools to handle other data and metadata formats Need some consistency to facilitate search and access across datasets Enable/Facilitate development of different interfaces to 	<ul style="list-style-type: none"> How do you build a system when the objects it is meant to process are evolving faster than the system can? How do you do that within the constrain of a budget cycle that is relatively rigid? While still meeting the business priority of the day? Coordination among 300+ user agencies Common 	<ul style="list-style-type: none"> Most commercial IT equipment has bit error rates of 10⁻¹⁴, including Ethernet backbone equipment: what good is storage BER of 10⁻¹⁷ when your system's best BER is 10⁻¹⁴ How often to check data integrity? <ul style="list-style-type: none"> Continuous above a certain size Reading the data can also damage it! How often to migrate? 	<ul style="list-style-type: none"> Maintaining an infrastructure to provide open and free to the user all of our science records is our biggest challenge. 	<ul style="list-style-type: none"> Creation, management, and preservation of digital collections Lack of a tangible collection from which to digitize publications Fugitive publications 	<ul style="list-style-type: none"> Asset Control Asset Sharing Misconception of DAMS as a 'Storage System' Explosion of digital assets exceeding SI staff resource bandwidth User Community education and awareness Potential for significant benefit to SI – currently only 25% of system functionality is being used. 	<p>Customers need improved weather services</p> <ul style="list-style-type: none"> US industry needs the most accurate, accessible, timely and reliable weather data to make critical decisions that impact our national economy Data explosion places demands on data management architecture <ul style="list-style-type: none"> Rapid data assimilation requirement 	<ul style="list-style-type: none"> What does it mean to unite systems? Ability to export data between systems <ul style="list-style-type: none"> Verify appropriate fixity Transparency for system administrators Ability to track collections between systems <ul style="list-style-type: none"> Verify collections are retrievable Verify collections retain original characteristics 	<p>The field of digital preservation is still emerging, and much is in flux—this mandates that we stay flexible and that we seek opportunities to collaborate with other solutions</p> <p>Cultural memory organizations face major decisions about their role as digital content stewards. Will they take their historical role seriously and continue to maintain</p>

		<p>support different communities</p> <ul style="list-style-type: none"> • Systems need to be designed to accommodate the evolution with time as technology changes rapidly • System changes over time need to be implemented and brought into operation without interruption in on-going services 	<p>agreement on security models and processes</p>	<p>- Individual files: every 5-10 years</p>				<p>s—aviation, onset of convection, forecast uncertainty Record retention - Data access on-demand within resource constraints</p> <ul style="list-style-type: none"> • Integrating all observing data sources to achieve desired effect and outcome 	<p>c</p> <ul style="list-style-type: none"> • What are the best ways to have an SRB/iRODS datagrid and a LOCKSS PLN interact? • What does it mean to have an active system (MetaArchive) and an archival system (Chronopolis) work together? • What are the appropriate transfer technologies? <p>- iRODS and LOCKSS native tools - CDL Micro-services, e.g. BagIt -</p>	<p>responsibility and control over digital assets in ways similar to their analog collections, or will they outsource this responsibility to other entities? If they choose the latter pathway, how can we guarantee that our digital cultural heritage will remain free and available to the public in ways that are consistent with our access to analog collections?</p>
<p>Lesson learned and future plans</p>	<ul style="list-style-type: none"> • Consensus on format across 40 or 60 databases often times the most difficult thing to achieve • Publishing the data in a least 	<p>Archive management:</p> <ul style="list-style-type: none"> • Ensure safe data stewardship through its lifetime • Perform regular peer reviews on archive holdings for scientific 	<ul style="list-style-type: none"> • The long term requirements for an Electronic Archives are leading us to an evolvable framework that need to support technology and needs 	<ul style="list-style-type: none"> • To reduce the number of physical carrier submissions by 50% • Less shelf space to pay for • Less air conditioning to pay for • Less cost for 	<p>Continue to publish Offline Archive Media Trade Studies to help us determine where those technologies are headed. Rely upon our appraisal process to ensure that we</p>				<ul style="list-style-type: none"> • Develop tools and methods to automate exchange of data between MetaArchive Cooperative (LOCKSS-based) and Chronopolis (iRODS- 	<p>Community-based organizing ensures that cultural memory organizations take a leading role in preservation activities. Collaboration brings costs</p>

	<p>common denominator format reduces richness inherent in high-quality databases</p> <ul style="list-style-type: none"> Database owners often reluctant to dedicate scarce resources for standardization 	<p>merit</p> <p>Data interoperability:</p> <ul style="list-style-type: none"> Enable multiple data and metadata streams to be seamlessly combined Enable interoperability between EOS and other research and value-added relevant data and systems Increase mobility of processing and data <p>Data access:</p> <ul style="list-style-type: none"> Make data location transparent and available with no delay Enable finding data via common search engines Increase services invoked by machine-machine interfaces Enable customizable data processing 	<p>that we don't even know about yet!</p> <ul style="list-style-type: none"> It is best that the Electronic Records Archive be built in such a way so as to fit in a technology ecosystem that can evolve naturally, and can be driven by the end users in ways that naturally ride the technology waves. The challenge is to co-exist and to leverage what's going on outside the Archival space. 	<p>Copyright holders if they don't have to pay for physical media</p> <ul style="list-style-type: none"> Less cost for us to maintain the equipment to ingest it into our system anyway! 	<p>are managing science collections that align to our mission.</p> <p>Attain a three-copy strategy for all of our electronic records with the last copy being off-site.</p>				<p>based)</p> <ul style="list-style-type: none"> Examine data transfer tools/protocols from: <ul style="list-style-type: none"> California Digital Library micro-services iRODS protocols for data transfer LOCKSS "plug-in" approach for data transfer Goal: A highly robust, easy to use preservation "system," allowing digital objects to be shared between several major preservation networks in the U.S 	<p>down.</p> <p>Partnerships with other digital preservation groups are extremely fruitful (and heterogeneous storage options appeal to Producers). Libraries and other cultural memory organizations can conduct their own preservation activities and ensure knowledge building and in-house expertise, especially when they work collaboratively to distribute content in secure networks.</p>
--	---	--	--	---	---	--	--	--	--	---

		<ul style="list-style-type: none"> • Universally employ open interfaces and best practice standard protocols <p>User support</p> <ul style="list-style-type: none"> • Ensure that expert knowledge is readily accessible to enable researchers to understand and use the data • Provide for direct community feedback to a given system element 								
Desire Tools	keep up with technologies – hardware upgrades, data migration, upgrade of software and tools to “keep up with the times”	<ul style="list-style-type: none"> • Evolvability & Extensibility • Scalability & Performance • Configurability • Ease of Use • Maintainability, Operability & Ease of Deployment 	<ul style="list-style-type: none"> • Direct data connections to major production centers • Transfer standardized file formats with standardized metadata • Automate the creation of database records and ingestion of metadata • Create files directly from all Congression 	<ul style="list-style-type: none"> • Evolvability & Extensibility • Scalability & Performance • Configurability • Ease of Use • Maintainability, Operability & Ease of Deployment 						<p>Better mechanisms for metadata extraction</p> <p>Better data exchange mechanisms between repository infrastructures</p>

				<p>al video and audio feeds (that's approx 5 PiB/year ALONE)</p> <ul style="list-style-type: none"> • Automaticall y QC everything 						
<p>Interoperability Needs</p>	<p>Yes , it will make connection to databases easier</p>	<ul style="list-style-type: none"> • Needed for different purposes and at different levels • Search and Access across systems: Directory , Inventory, Data levels • Not all systems need to interoperate – need is driven by user community requirements • Standards facilitate interoperability– difficult to “mandate” standards – easier to adopt community accepted standards 	<ul style="list-style-type: none"> • Need to provide access to records held in presidential libraries and partner websites. • Future need to provide system access to electronic records held by NARA to partner websites. • Need to ingest electronic records from external systems. 	<p>Yes, use standards as much as possible to enhance operation and management needs</p>	<ul style="list-style-type: none"> • Yes, use standards as much as possible to enhance operation and management needs 	<ul style="list-style-type: none"> • Use of standards in AIP • Use of API's • Support of Open Government Initiatives - FDsys Bulk repository contains XML content for download 			<ul style="list-style-type: none"> • Better definitions of standards for sharing data between diverse systems. Based on OAIS, what are the proper ways to move AIPS? 	<p>This is at the heart of the next 5 years of digital preservation work. We need ways of ensuring heterogeneous storage at acceptable costs. Moving content in and out of systems is the greatest challenge and expense in the digital preservation field right now. Better tools for interoperability (achieved through stronger standards that cultural memory organizations have the ability to implement) will help this enormously.</p>

<p>Willing to connect to and from other systems</p>	<p>Typically, databases were designed by people who had no reason to anticipate integrating their datasets with anyone else's</p>	<ul style="list-style-type: none"> • Yes, Core elements are working with Community elements for evolvability, innovation research, software reuse, and technology infusion • Open API's in EOS Clearing House middleware facilitates development of community-specific data search and access clients • Interoperability arrangements exist with international partners 	<p>Yes, system should be based on standards for import/export interfaces.</p>	<p>??</p>	<p>Yes, we have done this for many years and will continue to look for opportunities to link to and with systems or portals that provide additional services for our research community.</p>				<p>Yes, working with a number of systems to do this.</p>	<p>Absolutely. Working with Chronopolis (iRODS-based system) in this manner already.</p>
--	---	--	---	-----------	--	--	--	--	--	--

US DPIF Workshop: Technology

Submission Questions	Policy-Based Data Management (DICE/UNC)	Analyses of Electronic Records: A Framework for Understanding File Format Conversions (NCSA/UIUC)	Balancing Performance and Preservation: Lessons learned with HDF5 (HDF5)	Developing Bit Preservation Services at the Library of Congress (LOC)	Monitoring Distributed Collections Using the Audit Control Environment (ACE) (UMIACS)	Lap Around the Windows Azure Platform - Scalable Compute and Storage Cloud Environment (Microsoft)	Open Data Protocol (OData) - Querying and Updating Data on the Web (Microsoft)	DAITSS , an OAIS-based Preservation Repository (FCLA)
<p>Background Motivation</p>	<p>To provide policy-based preservation data life cycle management with properties of:</p> <ul style="list-style-type: none"> • Authenticity • Integrity • Chain of custody • Original arrangement • Trustworthiness <p>To automate validation of assessment criteria and execution of administrative functions</p>	<p>To support very large number of file formats in which digital content is stored, and by an increasing number of complex file formats containing multiple types of digital content (e.g., Adobe PDF, HDF) or having very elaborate specifications (e.g., STEP).</p>	<p>The HDF5 project aims to provide</p> <ul style="list-style-type: none"> • Outstanding technologies for managing large and complex data. • Excellent service to the users of HDF5. <p>Long-term availability and support for data stored using HDF.</p>	<p>As part of its first phase preservation tool development, the Library of Congress has been working on solutions for “Transfer.” Transfer includes the following human- and machine-performed tasks:</p> <ul style="list-style-type: none"> • Adding digital content to the collections, whether from an external partner or created at LC; • Moving digital content between storage systems (external and internal); • Review of digital files for fixity, quality and/or authoritative ss; and 	<p>To provide an infrastructure for managing integrity by securing items with small Integrity Tokens.</p> <ul style="list-style-type: none"> • Tokens allow all components to be externally validated • Scalable • Provide an externally auditable infrastructure 			<p>To develop a fully OAIS-conformant digital preservation repository system with these properties:</p> <ul style="list-style-type: none"> • can be used by multiple institutions • implements format-based preservation strategies • fully conformant with PREMIS and METS • maintains authenticity of content • facilitates use of existing tools, and provides tools that can be used by other applications

<p>Audience</p>	<ul style="list-style-type: none"> Archivists Digital Librarians Scientists sharing data Scientists implementing data processing pipelines Personal digital library 	<ul style="list-style-type: none"> managers of electronic records scientists conducting research with digital data, and citizens trying to keep their files current with the rapidly changing information technology 	<p>Virtually every individual or organization that must deal with complex or large scale data, including academia, government, and the private sector.</p>	<ul style="list-style-type: none"> Inventorizing and recording transfer life cycle events for digital files. <p>Any repositories interested in content transferred and preservation</p>	<ul style="list-style-type: none"> Managers of electronic records Any group or organization that needs to assert the integrity of its holdings to an outside party 			
<p>Technology Novelty</p>	<p>To provide policy-based data management:</p> <ul style="list-style-type: none"> Map from properties to policies to procedures to state information to queries that validate assessment criteria <p>To provide infrastructure independent, preservation environment for:</p> <ul style="list-style-type: none"> Middleware insulates records from changes in 	<p>Apply Conversion software Registry (CSR) which provides multi-hop conversion paths for file format conversions. As July 30th, 2010, it supports 1795 software packages representing 223,083 file format conversions.</p>	<p>The HDF5 data model is sufficiently flexible and general to accommodate virtually any kind and scale of data. The self-describing format of HDF5 can accommodate virtually any combination of data entities in a single package.</p>	<p>BagIT: A packaging specification for file transfers with self-identifying and self-describing along with support for error detection and transfer optimization. Motivation includes:</p> <ul style="list-style-type: none"> Low overhead Content-type agnostic 	<p>ACE provides a completely auditable environment which relies on no private information.</p> <ul style="list-style-type: none"> The ACE token allows integrity information to exist alongside data throughout its lifetime 			<p>Ensures long-term renderability by building into Ingest process format based preservation strategies forward migration and normalization.</p>

	<p>the external world</p> <ul style="list-style-type: none"> • Maps from procedures to new operating systems, protocols, and clients • Provides interoperability mechanism between old and new technologies 							
Architecture	<p>Client: allow users to search, access, add and manage data and metadata iRODS data system:</p> <ul style="list-style-type: none"> • Peer-to-peer servers • Data Server – disk, tape, etc • Rule Engine – track policies • Metadata Catalog – track information 	<p>Utilize computational cloud services to enable optimal and/or measurable data transformation from one data structure to another.</p>	<p>The HDF5 package consists of the format and a software library that implement the data model, command-line tools to help with data management, and HDFView, a tool for visualizing HDF5 files. The format and software combine I/O and storage efficiency, scalability, and platform independence in a single software package,</p>		<p>Integrity Management Service for token issuing, validation, and witness publication. Local Audit Managers monitor local files using token proofs for validation.</p>			<p>Series of RESTful Web Services.</p>
Areas of Application	<ul style="list-style-type: none"> • Data organization • Data processing pipelines 	<p>Services of interest includes:</p> <ul style="list-style-type: none"> • locate conversion 	<p>HDF5 is used to exchange, archive, and access data in</p>	<ul style="list-style-type: none"> • Transfer of content internally and between 	<p>Management of fixity information.</p>			<ul style="list-style-type: none"> • Long-term storage of content • Format based

	<ul style="list-style-type: none"> • Collection creation • Data sharing • Data publication • Data preservation • Personal collection • Federated data grids 	<p>software</p> <ul style="list-style-type: none"> • execute conversion process with any available third party software • provide information loss evaluation 	<p>every discipline, and many applications outside of science and engineering. Examples include NASA's Earth Observing System and related projects, simulation codes, tomography, DNA sequencing, flight and vehicle testing, product model data exchange, film-making, finance.</p>	<p>preservation partners.</p> <ul style="list-style-type: none"> • Long-term storage of content • Inventory system - Keeps track of and enables the querying of important events in the preservation lifecycle of a Bag and its contents 				<p>preservation strategies</p> <ul style="list-style-type: none"> • Migration • Normalization
Community	<ul style="list-style-type: none"> • NARA Transcontinental Persistent Archive Prototype • National Optical Astronomy Observatory • Carolina Digital Repository • NOAA/NCDC • NASA/NCCS • NSF/TeraGrid • French National Library • ARCS 	<p>Prototype services are open to public</p>	<p>See "areas of application." HDF has become a standard for many communities. An ISO standard is under development for product data representation and exchange in HDF5.</p>	<p>NDIIPP partners</p>	<ul style="list-style-type: none"> • Chronopolis consortium • GeoMapp 			<p>State University System of Florida (11 universities)</p> <p>DAITSS application is open source and available for use by anyone</p>

<p>Impact Benefits</p>	<ul style="list-style-type: none"> • Better federation with different vendor products • Better policy coalesce authentic records from independent data grids • Automation of administrative functions • Validation of assessment criteria • Policy evolution, versioning 	<p>Better understanding of preservation and reconstruction of electronic records in terms of file format conversions including:</p> <ul style="list-style-type: none"> • “optimal” file format to be preserved • Conversion software evaluation • Minimum cost on conversion path 	<p>HDF5’s ability to efficiently store almost any kind of data simplifies the packaging of data, which facilitates data integration, interoperability, and long term preservation.</p>	<ul style="list-style-type: none"> • Systematic method of transfer content • Self-identifying and self-describing package • Better inventory tracking and lifecycle package management 	<p>Better understanding of the integrity requirements through an objects lifecycle</p> <ul style="list-style-type: none"> • Best practices for monitoring data 			<p>Complete preservation solution for text, audio, video and image based formats.</p> <p>Web Services can be implemented any organization or consortium.</p> <p>Implementers can contribute to format based processing</p>
<p>Operation Enhancement</p>	<ul style="list-style-type: none"> • Easy generation of policy rules from chains of standard functions • Support recovery procedures in case of failures • Track status of all operations • Write results to persistent metadata catalog 	<ul style="list-style-type: none"> • Efficiency to manage file content holdings • Better understanding of information loss due to conversions 	<ul style="list-style-type: none"> • Efficient, scalable data storage and access. • Ability to keep all data in one container, facilitating interoperability and preservation. 	<ul style="list-style-type: none"> • Efficient and reliable method of content transfer • 	<ul style="list-style-type: none"> • ACE Audit Manager allows non-technical personal to manage data integrity 			<p>Best operated by a knowledgeable central staff.</p> <p>Each RESTful service is self-contained and relatively simple</p> <p>Detailed implementation and configuration instructions</p>
<p>Cost Saving</p>	<ul style="list-style-type: none"> • Supports interoperability from old to new technology • Open Source 	<p>Reduce conversion processing cost</p>	<p>Availability of capabilities that individual data management projects lack the resources or knowledge to</p>	<ul style="list-style-type: none"> • Eliminate duplication methods of content transfer protocols 				<p>Open Source</p>

			implement.					
Development Lessons Learned	<ul style="list-style-type: none"> • Preservation environment are inherently distributed and federated • Management of technology evolution is needed (migrate from old to new technology) • Preservation requires communication with the future and management of communication from the past • Federation minimizes risk of tampering • Periodic verification of assessment criteria is critical 	<ul style="list-style-type: none"> • Nobody has the resources to load every possible file format • Hard to support many available formats • Proprietary file format is hard to retrieve data • Vendor specific structures are not support from other formats <p>Conclusions: Software reuse and extensibility are the key characteristics of file format conversion systems</p>	<ul style="list-style-type: none"> • For long term preservation, simple models are better than complicated models, all else being equal. • Focus on being one good general purpose layer in the stack, letting other layers implement specialized capabilities. • Pay heed to backward and forward compatibility of both the format and the software. <p>Open source doesn't assure preservation.</p>		<ul style="list-style-type: none"> • It is not possible to continually audit large, offline collections. Other techniques (ie, sampling) are necessary • ACE Tokens are useful throughout the lifecycle of a digital object rather than just during preservation. 			<ul style="list-style-type: none"> • Scalability and reliability very important. Must run in high-volume production operation. • Original system written in 2006 was recently rearchitected from monolithic Java application to set of lighter weight RESTful Web Services written in Ruby – now easier to change, extend, reuse, and integrate external tools. • Implementing PREMIS elements and data model actually simplified, rather than complicated, development.
Future Plan	<ul style="list-style-type: none"> • Development of simple preservation environment interfaces • Preservation management 	Building 2 nd generation conversion services	Continue to build a strong foundation for sustainability, giving special attention to: <ul style="list-style-type: none"> • Standards - 	<ul style="list-style-type: none"> • Systematic inventorying of all Library content – allows the Library to track the location of 	<ul style="list-style-type: none"> • Release all components as open source software • Expand tools to allow integrity validation of cloud storage • Explore statistical 			<ul style="list-style-type: none"> • Implementation of re-architected version. • Add support for additional file formats.

<p>features including (a) format parsing routines and (b) representation metadata</p> <ul style="list-style-type: none"> • Automated creation of assessment policies • Development of standard preservation policy sets 		<p>develop domain-specific standards in use of HDF5, and pursue standardization of HDF5.</p> <ul style="list-style-type: none"> • Data model - work toward greater rigor in the data model. • Backward/forward compatibility - continue to develop processes and technologies for improving compatibility across generations of the formats and software. • Full support for HDF4 (earlier generation of HDF). • Preservation-based evolution for HDF5 - continue to actively develop the format and libraries to address current and future data 	<p>digital content and checksums to support auditing</p> <ul style="list-style-type: none"> • Development of additional workflows – expands other workflows to support on-going and future projects; better integration of UI to automate reliable, repeatable activities • Improved usability of the tools and services – through production usage of the tools and services, an ongoing iterative review and revision of interfaces is in place • Extension of our services from Bag-level to full file-level bit preservation • Support for the semantic mapping of files to objects and collections 	<p>sampling for large collections</p>			<ul style="list-style-type: none"> • Seek grant funding to support as Open Source project. • Promote more widely as a more fully featured alternative to storage-based solutions (e.g. data grids and private LOCKSS networks).
---	--	---	---	---------------------------------------	--	--	---

			<p>challenges</p> <ul style="list-style-type: none">• A sustainable institutional model - continue to explore organizational models that can improve sustainability of the institution and its technologies.					
--	--	--	--	--	--	--	--	--

Problem Areas	<p>Need for Digital Curation Education:</p> <ul style="list-style-type: none"> • “Mechanisms are also needed to accelerate the transfer of new knowledge into practical working digital preservation systems to prevent further loss of valuable digital collections.” From “It’s About Time” NSF/LC Report, 2003 • “The new discipline of digital preservation needs to be supported. This should include the provision of continual professional development for existing individuals with relevant skill sets, e.g., archivists, librarians and IT staff.” By Waller and Sharpe, “Mind the Gap: Assessing Digital Preservation in the UK,” DPC, 2006 	<ul style="list-style-type: none"> • • 		<p>Lack of advanced imaging standards Is undercutting broader scientific, medical, engineering, and commercial progress</p> <ul style="list-style-type: none"> • HPC • Interoperability • Archival 			<p>To develop a transfer format that supports the exchange of rich, heterogeneous archival information packages between heterogeneous repositories while maintaining an unbroken chain of digital provenance for the purpose of disaster recovery, succession planning, software migration, and/or specialized AIP processing.</p>
Audience	<ul style="list-style-type: none"> • individuals who want a master’s degree or our Digital Curation so 	<ul style="list-style-type: none"> • See Community under Standard 		<ul style="list-style-type: none"> • Electron microscopy • Science, math, medicine, 			<p>Preservation Repository managers and developers, those interested in</p>

	<p>as to be digital curators;</p> <ul style="list-style-type: none"> those people who want to earn a Ph.D. and be professors teaching about digital curation; People already in the field (Librarians, digital curators, digital asset managers, archivists, and museum collection managers) who need more skills in this area. 			engineering & commerce			<p>exchanging information between heterogeneous systems while maintaining an unbroken chain digital of provenance.</p>
Standard Novelty	<p>A Framework for Digital Curation Education</p> <ul style="list-style-type: none"> Through the DigCCurr projects we are building a framework for Digital Curation/ Preservation education. - Matrix of Digital Curation Knowledge and Competencies. - High-Level Categories of Digital Curation Functions. We are providing an environment in which to capture and education components that 	<p>PREMIS data model provides five entities particularly important to digital preservation community:</p> <ul style="list-style-type: none"> Intellectual Entity – a set of content or an intellectual unit (book, map, photograph, or database) for the purposes of management and description characteristics Object – a discrete unit of information in digital form Event – an action that involves or impacts at least 		<ul style="list-style-type: none"> Generation of file systems using HDF5 			<p>The Repository Exchange Package (RXP) is a hierarchical packaging format designed to facilitate the exchange of Archival Information Packages (AIPs) between digital repositories. The RXP employs METS and PREMIS metadata documents to encode AIP preservation and structural metadata at the root level. All other data are stored in the files subdirectory. Link to specification.</p>

	populate the framework: Digital Curation Exchange	<p>one Object or Agent associated with or known by the preservation repository.</p> <ul style="list-style-type: none">• Agent – person, organization, or software program/system associated with Events in the life of an Object, or with Rights attached to an Object.• Rights: assertions of one or more rights or permissions pertaining to an Object and/or Agent <p>With the above entities, PREMIS provides Information to help manage a resource for preservation purposes: about actions on an object (provenance)</p> <ul style="list-style-type: none">• Technical characteristics• Information about actions on an object (provenance)• Relationships (structural and derivative) <p>- Structural: indicates how compound objects</p>					
--	---	---	--	--	--	--	--

		<ul style="list-style-type: none"> - are put together - Derivative: results of common preservation actions • Rights metadata associated with preservation 					
<p>Architecture/ Workflow/ Best Practices Approach</p>	<p>DigCCurr I Project</p> <ul style="list-style-type: none"> • Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum. • A collaboration of the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill (UNC-CH) and the U.S. National Archives and Records Administration (NARA). Project ran July 1, 2006 – December 31, 2009. • Goals: <ul style="list-style-type: none"> - Build graduate-level curriculum. - Fund master’s student Fellows. - Build a network of digital curation educators and experts. <p>DigCCurr II Project</p> <ul style="list-style-type: none"> • Extending an 	<p>METS provides XML-based structure of digital objects and association with various kinds of metadata according to their components</p> <ul style="list-style-type: none"> • Uses XML schema to combine vocabularies from different namespaces for extensibility • Metadata can either be embedded or linked • Records names and locations of files the comprise either embedded or linked metadata objects • Provide hyperlinks between components <p>Associates executable behavior with components:</p>		<ul style="list-style-type: none"> • HDF5 • Files/filesystems hyperspaces 			<p>Repositories add a software module capable of reading and writing RXP. RXP can then be exchanged between any repository that can import and export RXP regardless of underlying repository architecture.</p>

International Digital Curation Curriculum to Doctoral Students and Practitioners.

- A collaboration of the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill (UNC-CH) and the U.S. National Archives and Records Administration (NARA).
- Project to run September 1, 2008 – August 31, 2012.
- Goals:
 - Continue to foster a growing and evolving network of international experts in the DC arena.
 - Develop a doctoral-level curricular framework; course content; and networked, distributed, International seminars to prepare future faculty to educate 21st century digital curators.

Areas of Application	<ul style="list-style-type: none"> • • 			<ul style="list-style-type: none"> • Long term storage of microscopy images • Advanced imaging • Advanced research 			Use cases: <ul style="list-style-type: none"> • succession planning • disaster recovery • software migration (repositories can easily export AIPs from an older repository implementation for ingest into a new repository implementation) • Inter-institution exchange for specialized AIP processing
Community	Closing the Digital Curation Gap Project: <ul style="list-style-type: none"> • Partners: UNC-SILS, IMLS, Joint Information Systems Committee (JISC) (UK), Digital Curation Centre (UK). • Focus on building tools for small and medium-sized cultural heritage institutions (LAMs). 	<ul style="list-style-type: none"> • PREMIS community • METS community • Digital repository community 		<ul style="list-style-type: none"> • EM, x-ray • Optical microscopy 			Preservation community and those interested in exchanging heterogeneous information packages between heterogeneous systems.
Impact Benefits	Educating Stewards of Public Information in the 21st Century (ESOPI-21): <ul style="list-style-type: none"> • ESOPI-21 seeks to prepare the next 	Benefits of using PREMIS in METS: <ul style="list-style-type: none"> • Packages together metadata necessary for digital preservation in a 		<ul style="list-style-type: none"> • Optimal computational performance • Optimal long-term maintenance • Advanced modeling 			Developed a packaging format that is relatively easy to implement and addresses use cases of interest to the Preservation

	<p>generation of public information stewards by building on the existing dual degree program offered jointly by UNCCH's School of Information and Library Science and its School of Government.</p> <ul style="list-style-type: none"> • Designed to prepare leaders in public information curation and public policy administration. <p>Closing the Digital Curation Gap (CDCG) Project</p> <ul style="list-style-type: none"> • The CDCG collaboration is designed to serve as a locus of interaction between those doing leading edge digital curation research, development, teaching, and training in academic and practitioner communities those with a professional interest in applying viable innovations within particular organizational 	<p>predictable format</p> <ul style="list-style-type: none"> • PREMIS provides technical and event metadata • METS provides structural metadata • Both standards are <ul style="list-style-type: none"> - Openly available - Flexible - Extensible - Maintained by an open process • Provides an exchange standard between repositories 		<ul style="list-style-type: none"> • Better data designs 			<p>community.</p>
--	---	--	--	---	--	--	-------------------

	contexts.						
Operation Enhancement	<ul style="list-style-type: none"> • • 			<ul style="list-style-type: none"> •Simplicity •Provenance •Optimal media transition over time 			Ability to exchange AIPs with any other repository capable of exporting and importing RXPs.
Cost Saving	<ul style="list-style-type: none"> • 			>\$1B			Easy migration from previous repository implementations. Provides a means by which repositories can implement disaster-recovery exchange agreements, e.g., Repository A and B agree to store 20 TB of each others AIPs.
Development Lessons Learned	<ul style="list-style-type: none"> • Content must be managed across the life cycle. • Life cycle extends well before and after repository life. • Not all repositories will be archival but archival principles, i.e., need to preserve authenticity of content, underlie all repositories. • Much of the difference between Digital Library work and Digital Curation has been focus – 	<ul style="list-style-type: none"> • Contents of each information package may vary depending on its function within a repository • Need to determine how to include representation metadata and associate it with package components • PREMIS data entities (objects, events, rights, agents) do not map perfectly to METS categories for representation metadata 		<ul style="list-style-type: none"> • Pixel performance is critical infrastructure • Interoperability is critical for interdisciplinary research & modeling 			<ul style="list-style-type: none"> • The effort required to create an RXP is quite reasonable. • Repositories may need to be modified to accept incoming digital provenance. • Maintaining a continuous record of ownership is tricky when an AIPs are transferred more than once. • Repositories need to analyze digital provenance identifiers to avoid namespace

	<p>access vs. preservation.</p> <ul style="list-style-type: none"> • Institutional repositories in academic libraries engage all the skills, requirements discussed here and are a great training ground. • Digital Curation is about maintaining and extending context over time and is essential for re-use. • Digital Curators must bridge between content creators/ producers and technologists. • Digital Curators need to anticipate future re-use of content. • Young and evolving field. 	<p>(techMD, digiProvMD, rightsMD, sourceMD)</p> <ul style="list-style-type: none"> • There are redundant elements between the two standards • Both have extensibility mechanisms • Flexibility of both standards requires implementation choices • Predictability will enhance the ability for exchange with minimal human intervention 					<p>clashes.</p> <ul style="list-style-type: none"> • Each use case carries its own demands for what must be maintained and/or understood by the receiving repository. • The RXP is only part of the infrastructure needed for inter-repository exchange. Repository managers will have to establish inter-repository agreements similar to service-level agreements spelling out requirements related to ownership, access, and preservation of received content.
<p>Future Plan</p>	<ul style="list-style-type: none"> • • 	<ul style="list-style-type: none"> • Changes to data model under discussion <ul style="list-style-type: none"> - Provide semantic units for intellectual entities - Make environment and significant properties/ characteristics separate entities 		<ul style="list-style-type: none"> • Consensus development • HDF FS development & operational integration 			<ul style="list-style-type: none"> • Continue testing and refining RXP specification. • Refine draft Inter-Repository Agreement • Continue TIPR dissemination activities

- | | | | | | | | |
|--|--|--|--|--|--|--|--|
| | | <ul style="list-style-type: none">• Revisions to XML schema<ul style="list-style-type: none">- Coordinate mechanism for extensibility with METS- Includes meta-metadata- Allows for more predictability for extensions• Extensible container for Agents | | | | | |
|--|--|--|--|--|--|--|--|

International DPIF Symposium: Content Organization

Submission Questions	NARA Electronic Records Archives (ERA) Lessons Learned and Future Direction (NARA)	Developing a Digital Preservation Programme at a National Library (New Zealand)	LOCKSS & LuKII Project (U. Berlin)	Digital Archives for Molecular Microscopy (EMBL)	Geo-Seas e-infrastructure (NERC)	The METAFOR project: preserving data through metadata standards for climate models and simulations (BADC)	The ESA Long Term Data Preservation Programme and Experiences Across CASPAR and GENESI-DR. (ESA)	Singapore National Library Technical and Operation Challenges and Future Direction (NLB)
Background Institution/ Department	<p>Study Earth from space to advance scientific understanding and meet societal needs via core and community data system elements :</p> <ul style="list-style-type: none"> Core – manage Earth science satellite mission data, airborne instrument data and field campaign/ in situ measurement data robustly to ensure continuity of research, access, and usability <p>Community – support data system evolution with technological innovations and develop data products to complement Core capabilities</p>	<ul style="list-style-type: none"> National Library of New Zealand The purpose of the National Library is to enrich the cultural and economic life of New Zealand and its interchanges with other nations by, as appropriate,— <p>(a) collecting, preserving, and protecting documents, particularly those relating to New Zealand, and making them accessible for all the people of New Zealand, in a manner consistent with their status as documentary heritage and taonga; and</p> <p>(b) supplementing and furthering the work of other libraries in New Zealand; and</p>	<p>Berlin School of Library and Information Science, Humboldt-Universität zu Berlin</p>	<p>- European Bioinformatics Institute, a unit of EMBL, the European Molecular Biology Laboratory (Intergovernmental organization)</p>	<p>The project is an Integrated Infrastructure Initiative of the EU FP7 Research Infrastructures Programme. The project includes 28 partner organisations from 17 maritime countries in Europe, including 26 marine geoscientific data centres</p>	<p>The main objective of METAFOR is to develop a Common Information Model (CIM) to describe climate data and the models that produce it in a standard way, and to ensure the wide adoption of the CIM. METAFOR will address the fragmentation and gaps in availability of metadata (data describing data) as well as duplication of information collection and problems of identifying, accessing or using climate data that are currently found in existing repositories.</p>	<p>The Earth Observation Directorate is focusing activities in the operational as well as in the scientific areas, both in space as in ground elements in order to increase Earth knowledge and support operational application, for the humankind benefits developing EO satellites and the related infrastructure:</p>	<p>NLB oversees both the National Library as well as the Public Libraries. By international convention, the functions of these two kinds of libraries are distinct and well-differentiated:</p> <ul style="list-style-type: none"> National Library Singapore (NLS): The National Library is every nation's knowledge institution, preserving its cultural and literary heritage as well as providing trusted reference services. Public Libraries Singapore (PLS): The Public Libraries Singapore (PLS) provides a professional and engaging public library service to Singaporeans in their pursuit of lifelong learning

		<p>(c) working collaboratively with other institutions having similar purposes, including those forming part of the international library community.</p> <ul style="list-style-type: none"> • 						and discovery through the network of 22 Public Libraries (including three regional libraries) located conveniently across Singapore.
Offering	<p>Provide end-to-end capabilities to deliver Earth science data and information products to users</p> <p>Provide “one-stop-shopping” search and access to datasets in distributed Data Centers based on specific Earth science disciplines, as well as tailored access to datasets individual Data Centers.</p>	<p>Establish the National Digital Heritage Archive to enable the National Library of New Zealand to meet its mandate to collect, make accessible, and preserve in perpetuity, New Zealand’s digital heritage.</p>	<p>Interoperability between digital archiving systems LOCKSS and kopal.</p>	<ul style="list-style-type: none"> • Curated community database for high-resolution electron microscopy images of macromolecular complexes and subcellular structures 	<p>A pan-European e-infrastructure for marine geoscientific data, data products and data services, allowing federated access to data and products via the Geo-Seas data portal.</p>	<p>The Common Information Model (CIM) in the form of a conceptual model defined in UML, and an application model in xml, along with tools and services to take advantage of the CIM’s ability to categorize and relate the metadata of climate models.</p>	<p>Provide capabilities to deliver earth Observation data basic products to researchers and commercial customers</p>	<ul style="list-style-type: none"> • Provide a trusted, accessible and globally-connected library and information service so as to promote a knowledgeable and engaged society.
Audience	<p>Scientific and applications users, general public, other federal agencies, educators, international partners, value-added service providers.</p>	<ul style="list-style-type: none"> • General public • Researchers • Students • etc 	<ul style="list-style-type: none"> • Researchers & libraries in Germany 	<ul style="list-style-type: none"> • Structural biologists; biomedical researchers 	<p>Primarily scientific, economic, planning and environmental management communities within Europe, but can be accessed globally.</p>	<p>Climate modelers and those who will make use of climate models, such as policy makers and scientists in the climate impacts and adaptation communities.</p>	<ul style="list-style-type: none"> • Commercial applications • Scientific community • Education • Climate change • Disasters monitoring • Agriculture applications • Urban growth • Security • International communities • International initiatives (GEO, 	<ul style="list-style-type: none"> • General public • Researchers • Students • etc

							CEOS, UN, etc.) <ul style="list-style-type: none"> • Public sectors • Media • Etc. 	
Impact	Benefits	<ul style="list-style-type: none"> • Enabling climate modeling, global change analyses, disaster response and impact analyses • Supporting near real-time applications in many societal benefit areas (agricultural efficiency, air quality, aviation safety, carbon management, coastal management, ecosystems, disaster preparedness, energy forecasting, homeland security, invasive species, public health, and water management.) • Maintaining long-term., consistent Earth science data records relevant to climate and other terrestrial phenomena • 	<ul style="list-style-type: none"> • Preservation of New Zealand's digital heritage • Reduce duplication of effort in culture and heritage sector and government • Leverage cost savings through provision of 3rd party services 	<ul style="list-style-type: none"> • Combines the best features of two major archiving systems. 	<ul style="list-style-type: none"> • Distribution of research results for further use • Enabling peer-review and validation 	<p>Marine geological and geophysical data are very expensive to acquire, and it can take a few years to organise and conduct a new offshore survey if one is required. So there are both economic and time-saving reasons for the re-use of existing data where possible. Geo-Seas will allow users to locate, assess the quality of and acquire harmonized and federated marine geoscientific data and data products through the Geo-Seas data portal.</p>	<p>Tools will allow easier discovery and comparisons of climate model metadata and data. The controlled vocabulary developed for the CIM will be of use to the wider community, even without the use of the CIM.</p>	<ul style="list-style-type: none"> • Increase climate knowledge, modeling, global change analyses, disaster response and impact analyses • Others...

<p>Consequence if content not preserve</p>	<ul style="list-style-type: none"> Loss of data would have a negative impact on future verifiability of conclusions from global change analyses Data from NASA missions are records of physical phenomena measured at specific times and, if lost, can never be recreated. 	<p>Loss of New Zealand's digital heritage</p>	<p>Open access research material not preserved in publisher houses would be lost.</p>	<ul style="list-style-type: none"> Reduced research opportunities Duplication of effort Inability to verify published results 	<ul style="list-style-type: none"> Difficulty in locating and accessing data held by many different data centres in many countries, which may have different metadata standards, nomenclatures formats and levels of data aggregation. Increased costs of data acquisition or increased uncertainty of interpretations and models if data not available 	<p>If the metadata of the climate models is not preserved, crucial information about the model results will be lost, affecting reproducibility and reanalysis of the data – not to mention the political aspects of transparency and accountability of climate model results.</p>	<p>Data preservation and data access are of paramount value to avoid gaps in future known and new application in several fields, including Earth and atmosphere global change</p>	
<p>Benefits of using stds DP</p>	<ul style="list-style-type: none"> Enhance operation and management of data <p>Improve communication / interface between internal and external systems</p>	<ul style="list-style-type: none"> Enhanced global access to digital resources over time Ability to leverage other institution's experiences Enhanced operational management locally and internationally Enhanced interoperability between preservation services 	<p>Using LuKII offers a well-tested solution (LOCKSS) that incorporates standard migration and metadata methods (kopal).</p>	<ul style="list-style-type: none"> 	<ul style="list-style-type: none"> Access to data, data products and data services from the project data centres, via the Geo-Seas data portal, will have a significant impact on the time and resources required by users to search for, enquire about, order, acquire, process and merge data held by individual data centres across Europe. This will encourage more re-use of the data by both specialists and 	<p>The CIM is an effort to create a metadata standard for the climate modeling community.</p>	<ul style="list-style-type: none"> Ensure and increase operation and data exploitation Enhance the end to end systems from data gathering and archival up to the end users, for faster and improved services Improve the overall quality and homogeneity of products Aim at standardization of interfaces and interoperability of archives 	

					non-specialists alike.			
Content	<p>Size</p> <ul style="list-style-type: none"> • 4.2PB total archive volume • 1.8TB daily growth • >4000 unique data products • >910K distinct users of EOSDIS data and services • > 1M web site visits of 1 minutes or more • >254M end user distribution products • 6.7TB/day of end user average daily distribution volume • (All of the above are metrics for FY 2009) 	<p>Current programme 128TB</p> <p>Legal deposit, web archiving (including .nz domain crawls), digitisation, sound preservation</p> <p>Newspapers:</p> <ul style="list-style-type: none"> • 61 titles • 289,589 issues • 1,634,757 pages • 19,588,060 articles 	N/A	<ul style="list-style-type: none"> • 877 entries as of August 2010 • nearly 200 new depositions per year • 28 Gbyte total data size 	<p>The volume of data (digital and analogue) managed by the data centres is currently unknown. This will become clearer once the Geo-Seas metadata catalogue is compiled. However, the volume of digital data will probably be in the hundreds of terabytes. The bulk of this will be raw survey data with a lesser volume of processed data.</p>	<p>CIM portal and database is not yet operational.</p>	<ul style="list-style-type: none"> • 4.2PB total archive volume • 1 TB daily growth • About 200 data products and services • About 20000 users of EO data and services • 1 TB/day average daily distribution to users 	<ul style="list-style-type: none"> • National Libraries focus on preserving the Cultural Heritage - Digital Heritage : Consists of unique resources of human knowledge and expression (UNESCO, Paris 2003) - Covers cultural, educational, scientific, administrative resources, technical, medical, and other kinds of information created digitally or converted into digital form. • Audio Archive - Music Library in collaboration with Archive of Contemporary Music

<p>Scope</p>	<p>Data derived from imagers, sounders, active and passive sensors on board satellites; airborne instruments and in situ measurements</p> <p>Data from US and international instruments on board US and international partners' satellites</p> <p>Raw data and derived digital scientific "standard" and near-real-time data products covering many Earth science disciplines – atmospheric, oceanographic, cryospheric, seismic, land processes and solar radiance</p>	<p>All published and unpublished digital output of the nation including some territories in the Pacific. Actual ingest determined by relevant collection/selection policies.</p>	<ul style="list-style-type: none"> • All digital archiving in Germany. 	<ul style="list-style-type: none"> • Public distribution 	<p>Marine geoscientific data held by European marine data centres.</p>		<p>The end to end infrastructure must be adequate and properly operated in order to fulfill the mandate and the objectives of ESA , as they are committed to the funding ESA Member States, in order to be in line with the missions commitment and the long term data preservation strategy</p> <p>A Long Term Data Preservation Program (LTDP) has been approved recently by the ESA Member States in order to implement in long terms a European LTDP framework involving possibly all other European Earth Observation data owners /providers</p>	
<p>Distribution mechanism</p>	<ul style="list-style-type: none"> • On-line protocols (ftp, http, etc.) and services 	<ul style="list-style-type: none"> • Online applications • Aggregated metadata services • Google etc 	<p>LOCKSS alliance</p>	<ul style="list-style-type: none"> • Web-based submission and retrieval • FTP archive 	<p>Following a search of the central data catalogue – the Common Data Index (CDI), orders for data and data products will be made online via the portal using a shopping-basket style interface. Once the data are ready, the user is notified that the data are ready for download. Note that for extremely</p>	<p>CIM documents will be distributed via atom feeds</p>	<ul style="list-style-type: none"> • On-line protocols (ftp, http, etc.) and services • Physical media delivery 	

<p>Technologies (metadata, file format, packaging, etc.)</p>	<p>Metadata:</p> <ul style="list-style-type: none"> • ECS (EOSDIS Core System) metadata • FGDC-STD-012-2002 metadata for remote sensing data • EOS Clearing House (ECHO) metadata • ISO 19115 for geographic information and services metadata <p>File format:</p> <ul style="list-style-type: none"> • Primary HDF-EOS • HDF4 and HDF5 • NetCDF • GeoTIFF • CEOS • Binary <p>Others...</p>	<ul style="list-style-type: none"> • Dublin Core • Marc • OAI • PREMIS • XML • SRU/SRW • METS 	<ul style="list-style-type: none"> • Uses METS and WARC files. 	<ul style="list-style-type: none"> • XML for metadata • Community data formats for images 	<p>large data sets, delivery will be on tape or disk media.</p> <p>Uses ISO 19115, ISO 19139, OGC standards plus extensions. Conforms with INSPIRE. SOAP web service technology is used to connect the data centres. A small set of international, data-specific file formats (for example, SEG-Y and NetCDF) have been agreed and used by all partners for data delivery.</p>	<p>Xml and xsd schemas. UML</p>	<p>Metadata:</p> <ul style="list-style-type: none"> • Centralized catalogue with browsing capabilities • Interactive search of products availability based on temporal/geographical/features parameters <p>File format:</p> <ul style="list-style-type: none"> • HDF • CEOS • GeoTIFF • SAFE format • Missions specific • Others... 	<ul style="list-style-type: none"> •
<p>System Architecture</p>	<ul style="list-style-type: none"> • Geographically distributed (across the U.S.) system of systems • Scalable commodity hardware • Combination of custom and off-the-shelf software <p>Well-defined inter-system interfaces and API's</p>	<p>IBM XIV tiered storage (disk and cache copy), IBM TS3500 tape silo (3 copies), Oracle database, Rosetta digital preservation application, a range of web based resource discovery applications.</p>	<p>Linux</p>	<ul style="list-style-type: none"> • Web-based • Java and Python 	<p>The original data and metadata are held by the individual data centres using their own internal data management systems. CDI records are generated locally as CDI XML schema files and exchanged with the central data portal as XML files. The central Request Status Manager (RSM) contacts the local Download Manager clients at each site, which then</p>	<p>To be determined</p>	<ul style="list-style-type: none"> • Hierarchical archive infrastructure based on SAN technology Tape library • 13 distributed archiving facilities • StorageTek SL 8500 tape libraries based on T10000 tapes with 1TB/tape • 20 PB licence available • Current availability of up to 10 PB • Planned evolution of 	

					packages the data at each site ready for download or physical delivery to the user.		technology every 5 years aiming at 40 PB capacity by 2020	
OAIS relevant	Yes. Parts of the system follow OAIS reference model. Examining other parts for compliance and/or functional mapping	Rosetta has multiple functions based on the OAIS reference model which provided the base template for the development of the Library's digital preservation programme.	Yes	Not in use	Yes		Fully compliant with the entire reference model of OAIS	
Access								
Methods	<ul style="list-style-type: none"> • Most data are available via ftp • Some data in near-line robotic tape on-line services for searching, subsetting, reprojection, mosaicing, format conversion, statistical analyses, visualization etc. 	<ul style="list-style-type: none"> • Web based metadata aggregations (eg find.natlib.govt.nz), specialist web based applications (eg paperspast.natlib.govt.nz for newspapers). 	HTTP / OAI-PMH, & others.	<ul style="list-style-type: none"> • Web/HTTP • FTP 	Access to the data catalogues, data and data products is via the Geo-Seas data portal (www.geo-seas.eu).	Creation of CIM documents will be done via a web-based questionnaire. Automatic generation from other metadata standards is being considered	<ul style="list-style-type: none"> • All data accessible via internal procedures from tape libraries in near line • 20 % of data available via ftp in real time • Plan to reach 50% in few years using large disc as front end • Final products Rolling archives for some missions • Data are stored mostly in raw form and need to be processed for users usability. 	
Restrictions	All data are available at no cost to all users except where agreed upon with international partners	Copyright, donor usage restrictions, etc are respected as appropriate. Primary principle is that government information should be publicly available.		<ul style="list-style-type: none"> • None 	Access to the catalogue is open. There may be some restrictions on access to some data types depending on organizational or national licensing requirements, but	Currently the CIM only deals with global climate models. This will be extended to regional climate models.	All data are available at no cost for scientific application Currently some missions data delivery is based on commercial distribution Data distribution	

					the aim is for open access. Any restrictions will be described by the catalogue.		policy currently under revision in order to have all data free distribution	
Management What content to preserve	<ul style="list-style-type: none"> • Observation data from mission instrument (raw data, derived key data products, geo radiance data, etc.) • Validation field campaign datasets and inter-comparisons with other instruments • Other agencies' ancillary datasets • Results from derived high-level products, applications and research • Pre-flight or preoperational performance measurements • Instrument / sensor calibration data and method, etc. 	<ul style="list-style-type: none"> • See scope. Legislative mandate defined by relevant collection and selection policies. 	Any	<ul style="list-style-type: none"> • All deposited data will be preserved 	Examples of the types of geological and geophysical data are raw observational and analytical data, as well as the derived and processed data products from visual observations, seabed sediment samples, boreholes, geophysical surveys (seismic, gravity etc), side-scan sonar surveys and multibeam surveys.	CIM xml documents	<ul style="list-style-type: none"> • All EO operated satellite transmitted data instruments in raw form, ancillary data, auxiliary data, including algorithms documentation • Calibration data datasets and inter-comparisons with other instruments 	
Duration	Irrecoverable data need to be preserved indefinitely. Processes need to exist for determining duration of preservation for other data, derived products, etc.	Our legislation requires 'in perpetuity'.	Unlimited	<ul style="list-style-type: none"> • Perpetual archive 	The data are mainly from the 1960s onwards, and new data are being acquired continually.	Indefinitely	In principle forever	

Lifecycle strategies	<ul style="list-style-type: none"> Periodically refresh media including 'touching' all data Budget for hardware refresh every three years Metadata is a key cost driver <ul style="list-style-type: none"> needs to be continually reconciled and updated changes with each new data model websites are useless without good metadata <p>Science discipline expertise is required for management of data</p>	<ul style="list-style-type: none"> Media refreshment, migration, emulation. Risk management is primarily based on formats and our ability to render formats within our organisation. Ongoing referential integrity checks between database, METS/XML representations of objects and actual objects in the file system. 	<p>Migration, emulation, and (when needed) digital forensics.</p>	<ul style="list-style-type: none"> Integration in EBI bioinformatics data strategy 	<p>Each data centre has its own strategy depending on its resources and policies.</p>	<ul style="list-style-type: none"> Unknown as yet 	<ul style="list-style-type: none"> Periodical migration of media to new technologies Maximum five years technology cycle Catalogue updates in case of data loss or quality purging Evolution applies to the entire infrastructure following cost analysis (investment, maintenance, performances, etc.) and budget availability 	<ul style="list-style-type: none"> Acquire, describe, store, access, and preserve
Standards Used	<p>OPeNDAP, OGC standards where they apply Also see Technologies under Content</p>	<p>Open standards wherever possible. Proprietary vs open source is determined by fit for purpose. Digital preservation or content management will be our core business activity within 5-10 years and needs to be managed with enterprise class systems.</p>	<p>Many</p>	<ul style="list-style-type: none"> Community standards for file formats XML ad-hoc 	<p>Each data centre has its own standards although and aim of the project is to try to harmonize data management standards across the data centres.</p>		<p>CCSDS OAIS ISO Archives certification applicable standards</p>	
Protection (Sensitivity (Privacy Confidentiality Security Intellectual Property))	<ul style="list-style-type: none"> None for scientific data and derived products with the exception of those governed 	<p>All IP positions respected (eg copyright, donor usage restrictions etc). Primary principle is that</p>	<ul style="list-style-type: none"> All copyright protections are honored. 	<ul style="list-style-type: none"> Unreleased entries kept confidential Released entries completely public 	<p>The aim is to develop a common data policy and licence protocol for accessing data and copyright issues.</p>		<p>Security measures implemented in order to avoid facilities intrusion, computer hacking, sensitive data</p>	

	<p>by international agreements</p> <ul style="list-style-type: none"> • User-related information is protected in compliance with privacy regulations • Some documentation is protected as SBU or ITAR-sensitive 	<p>government information should be publicly available. These issues are managed at the infrastructure and Rosetta application level.</p>			<p>However, because the data centres have their own internal and possibly national regulations, there may be restricted access to certain data held by individual data centres. Any restrictions will be highlighted in the catalogue record, and the individual data centre contacted to negotiate any licence or approvals required to access data.</p>		<p>protection procedures, data delivery confidentiality, data integrity with archives duplication, disasters protection</p>	
<p>Challenges Issues</p>	<ul style="list-style-type: none"> • Scientific disciplines have different ways of looking at the data and different vocabularies. • Need flexibility and tools to handle other data and metadata formats • Need some consistency to facilitate search and access across datasets • Enable/Facilitate development of different interfaces to support different communities • Systems need to be designed to 	<ul style="list-style-type: none"> • Need enterprise class tools. Current tools not satisfactory (eg DROID, JHOVE, NLNZ Metadata Extract Tool) • Need enterprise class services for preservation strategies (migration, emulation) that also provide verifiable quality assurance mechanisms • Lack of agreed definitions for digital preservation systems • Lack of skilled resources in digital preservation, 	<ul style="list-style-type: none"> • Integrating systems with different design strategies. 	<ul style="list-style-type: none"> • Ensuring deposition of data; reach out to community • Technical hurdles (incompatibilities, data size, changing software environments) • Metadata acquisition (currently manual, in future integration with other software) 	<p>Creating the environment (legal framework and finance) to build the consortium and fund the development.</p>	<ul style="list-style-type: none"> • 	<ul style="list-style-type: none"> • Increasing demand of data in terms of volume , data delivery speed, products homogeneity, large datasets reprocessing • Data mining search • Interoperability between different data providers • Products format standardization • Exponentially increase amount of data to acquire, store and process, ensuring real time access • Standardize interfaces to 	<ul style="list-style-type: none"> •

	<p>accommodate evolution with time as technology changes rapidly</p> <ul style="list-style-type: none"> • System changes over time need to be implemented and brought into operation without interruption in on-going services 	<p>both at the micro level (eg managing formats, bit rot) and the macro level of specialist data areas (astronomy, oceanography etc)</p> <ul style="list-style-type: none"> • Lack of strategic understanding of the increasing pace and complexity of digital change across all disciplines going forwards • A meaningful interoperability framework for global digital preservation systems • Need for a global model of digital preservation connecting sectors, disciplines etc 					support different communities	
Lesson learned and future plans	<p>Archive management:</p> <ul style="list-style-type: none"> • Ensure safe data stewardship through its lifetime • Perform regular peer reviews on archive holdings for scientific merit <p>Data interoperability:</p> <ul style="list-style-type: none"> • Enable multiple data and metadata streams to be 	<ul style="list-style-type: none"> • Future plans relate primarily to challenges above and growing our capability and capacity in digital preservation. Cross sectoral collaboration is going to be key, reaching out of our own discipline to find the commonalities for digital 	<ul style="list-style-type: none"> • Expand interoperability to other systems. 	<ul style="list-style-type: none"> • Integration with the community is crucial to ensure depositions (advisory committees and task forces, attendance at conferences) • Technical competence through integration with similar efforts in other fields (e.g. light 	<p>There is a general willingness at the data management level in all organisations to cooperate in making data available using common standards. The challenge is putting the framework in place and persuading funding agencies that data management is important. This is</p>	<p>Important to seek community input from the beginning of the project.</p> <p>Maintaining a clear distinction between conceptual and application schema has been very helpful.</p> <p>Future plans: convert the CIM to GML compatible framework.</p>	<ul style="list-style-type: none"> • Aim at a multi-mission clear mandate for data archive management • Define a consistent data preservation policy based on an an hoc assigned budget in long terms • Careful and regular migration of data to new technologies • Aim at fully 	

<p>seamlessly combined</p> <ul style="list-style-type: none"> • Enable interoperability between EOS and other research and value-added relevant data and systems • Increase mobility of processing and data <p>Data access:</p> <ul style="list-style-type: none"> • Make data location transparent and available with no delay • Enable finding data via common search engines • Increase services invoked by machine-machine interfaces • Enable customizable data processing • Universally employ open interfaces and best practice standard protocols <p>User support</p> <ul style="list-style-type: none"> • Ensure that expert knowledge is readily accessible to enable researchers to understand and use the data • Provide for direct community feedback to a given system 	<p>preservation within 'apparently' disparate practices. This should help us towards more seamless, user friendly access systems for the full range of user communities, in a manner consistent with the needs of those communities.</p>		<p>microscopy, semantic web)</p> <ul style="list-style-type: none"> • Database should try to add value through <ul style="list-style-type: none"> • validation/quality assessment • leveraging integration/linking with other databases 	<p>even more complicated for international projects.</p> <p>The current project lasts for 3.5 years. The aim is to seek longer-term funding nationally and from the EU.</p>		<p>automatic operations</p> <ul style="list-style-type: none"> • archived Data and production facilities to be fully transparent to users • Data search via similar tools and procedures • Support standard protocols • Be more user requirements oriented in the system design and not only technology driven • Planned by 2012 the integration of the 13 facilities based on an internal GRID concept, where different locations become virtual and data flow is totally transparent to the external access,, in order to ensure performances, security, and reach the maximum exploitation of the IT infrastructure without down time and services interruptions 	
---	--	--	---	---	--	--	--

	element							
Desire Tools	<p>keep up with technologies – hardware upgrades, data migration, upgrade of software and tools to “keep up with the times”</p> <ul style="list-style-type: none"> • Need enterprise class tools. Current tools not satisfactory (eg DROID, JHOVE, NLNZ Metadata Extract Tool) • Need enterprise class services for preservation strategies (migration, emulation) that also provide verifiable quality assurance mechanisms 	N/A	<ul style="list-style-type: none"> • Metadata management (semantic web, ontologies) • Integration over different databases • Integration with emerging scientific data management systems 	None			<p>The procurement of the most adequate technology that represents the best compromise between operational requirements and budget limitations</p>	
Interoperability Needs	<ul style="list-style-type: none"> • Needed for different purposes and at different levels • Search and Access across systems: Directory , Inventory, Data levels • Not all systems need to interoperate – need is driven by user community requirements <p>Standards facilitate interoperability– difficult to</p>	<ul style="list-style-type: none"> • Imperative for a global approach to digital preservation • Standardisation • Agreed definitions for digital preservation systems • Agreed metadata for digital preservation • Focus on digital preservation, let collection management, resource discovery take 	Testing interoperability	<ul style="list-style-type: none"> • Deposition: interface with user software • Retrieval: interface with other databases • On-line visualization (e.g. web services, weblets) 	None identified	None – using standard methods of transferring xml files (atom feeds)	<ul style="list-style-type: none"> • Search and access across different inventories / archives of several EO products providers, ensuring single shop stop virtuality • Need of commonly adopted standards and the implementation of the necessary interfaces 	

	“mandate” standards – easier to adopt community accepted standards	care of themselves						
Willing to connect to and from other systems	<ul style="list-style-type: none"> • Yes, Core elements are working with Community elements for evolvability, innovation research, software reuse, and technology infusion • Open API's in EOS Clearing House middleware facilitates development of community-specific data search and access clients <p>Interoperability arrangements exist with international partners</p>	Yes	Yes	Yes, always.	Yes	Yes	All internal and external systems to be able to dialogue as part of a coordinated orchestration	

International DPIF Symposium: Technology

Submission	Policy-based Data Management (DICE/UNC)	Quality Assurance: Towards Tools for Characterizing and Comparing Digital Documents (Microsoft UK)	The Planets IF - A Framework for Integrated Access to Preservation Tools (Planets)	The eXtensible Characterization Languages – XCL (U. Koln)	100 Million Hours of Audiovisual Content: Digital Preservation and Access in the PrestoPRIME Project (IT Innovation)	EuroVO Framework and Future AIDA Direction (SADC)	CASPAR Framework and Lesson Learned (STFC)	PARSE Insight Framework and Lesson Learned (STFC)
Questions								

<p>Background Motivation</p>			<ul style="list-style-type: none"> • Planets is a four year integrated project funded by the European Commission • The project is driven by requirements for long-term preservation faced by institutional libraries and archives. • The Planets Interoperability Framework (IF) provides the technical backbone for integrating existing content repositories, preservation tools, and services into a homogeneous infrastructure. 				<p>EU FP6 Integrated Project with total spend of 16M Euros; 17 partners led by STFC. The aim was to investigate and develop digital preservation techniques, providing evidence the effectiveness of preservation strategies. Produce prototype implementations of preservation components. Identify ways of improving preservation capabilities of existing archives.</p>	<p>EU FP7 Support Action with 9 partners led by STFC. The aim was to produce an evidence based Roadmap for digital preservation infrastructures. The evidence was collected from surveys and case studies from researchers, data curators, publishers and funders.</p>
<p>Audience</p>		<ul style="list-style-type: none"> • 	<ul style="list-style-type: none"> • Digital Preservation Research Community • Libraries, Archives • Computer Science - Distributed and Service-oriented Architectures 				<p>Researchers in digital preservation, practitioners of digital preservation, funders of digital preservation.</p>	<p>Everyone concerned with digital preservation.</p>

Technology Novelty	•		<ul style="list-style-type: none"> • Web service technology, On-demand Computing, Repository Systems, Legacy System Integration • Provides defined interfaces for preservation actions • Uniform access mechanisms to a broad range of commodity tools and repository systems. • Workflow execution, provenance, preservation metadata generation 				<p>Implementation of OAIS concepts, in particular the Information Representation Model including Representation Information from the bits upwards, tested against many types of data.</p> <p>Application of Knowledge management techniques to digital preservation including Semantic Rep. Info., preservation workflows, definition of a Designated Community..</p> <p>Create a "preservable" infrastructure for preservation through technology neutral design.</p>	No s/w developed.

Architecture	•		<ul style="list-style-type: none"> The architecture is SOA based and comprises of three tiers: Portal representation, gateway server, service layer. 				Extremely distributed, heterogeneous, asynchronous architecture with no single point of failure and multiple deployment strategies.	N/A
Areas of Application	•		<ul style="list-style-type: none"> Development and evaluation of preservation tools and strategies for human-centric data holdings 				All types of digitally encoded information. Tested using data from science, cultural heritage and contemporary performing arts.	All aspects of digital preservation.
Community			<ul style="list-style-type: none"> Digital Preservation, Repositories, e-Science 				All communities	All communities

<p>Impact Benefits</p>	<ul style="list-style-type: none"> • 	<ul style="list-style-type: none"> • 	<ul style="list-style-type: none"> • Evaluation and application through various memory national institutions • The IF is implicitly utilized through the Planets applications and their users. • Sustainability through Open Planets Foundation • Open-Source code available through SourceForge 				<p>Evidence based support of claims about digital preservation which allows archives to use as much or as little as they wish to improve their preservation capabilities. Allows sharing of effort in digital preservation across all communities.</p>	<p>Evidence base for decisions about key attitudes and appreciation of threats about digital preservation.</p>
<p>Operation Enhancement</p>	<ul style="list-style-type: none"> • 	<ul style="list-style-type: none"> • 	<ul style="list-style-type: none"> • Easy and seamless access to broad and extensible range of preservation tools, independently from language or platform • Integrated access to existing repository and storage facilities • Automation and batch execution through integrated workflow engine 				<p>Many preservation aspects of any repository can be enhanced with proven preservation tools and techniques.</p>	<p>Information about digital preservation to help in decision making.</p>

Cost Saving	•		<ul style="list-style-type: none">• Provides a distributed environment allowing multi-user and cross-organizational deployments.• Sharing of hardware, software, maintenance• Easy and uniform access methods hiding complexity of tools and architecture• Provides back-end infrastructure for automated decision support and benchmarking applications.				Design facilitates sharing of information and development of Representation Information – which is absolutely critical to preservation.	Not directly applicable.
--------------------	---	--	--	--	--	--	---	--------------------------

<p>Development Lessons Learned</p>	<ul style="list-style-type: none"> • 		<ul style="list-style-type: none"> • A need for defining atomic preservation actions in order to compose more complex strategies. • Automated Quality control is a big issues • Interoperability required on semantical and technical level • Integration of 3rd party components, legacy code, and proprietary systems causes drawbacks • Missing approaches for the application of preservation actions to preservation repositories, in particular with respect to scalability and trustworthiness 				<p>It is difficult to explain some of the OAIS concepts to practitioners who do not have experience with data. IN particular the importance of Semantics and Semantic Representation Information can be ignored when dealing with rendered objects but not for data.</p>	<p>Many interesting lessons about attitudes and ideas about digital preservation.</p>
<p>Future Plan</p>	<ul style="list-style-type: none"> • 		<ul style="list-style-type: none"> • Looking into robustness, scalability and automated quality control 				<p>A clear distinction was made in CASPAR between components which are applicable to any type of data and those tools which were domain specific. The plans are to</p>	<p>The Roadmap defines components for an infrastructure to support digital preservation. These are consistent with the CASPAR components.</p>

								improve the scalability and robustness of the former and make them part of the preservation infrastructure. The domain specific tools will be developed further and new tools created.	Future plans include implementing this Roadmap.
--	--	--	--	--	--	--	--	--	---

International DPIF Symposium: Standards and Best Practices

Submission Questions		***** The Usage of MPEG-21 Digital Items in Research and Practice (U. Klagenfurt)	Digital Preservation: The Multimedia Standards way (U. Passau)	***** Introduction to MPEG-A Professional Archival Application Format (PA-AF) (NTT)	***** Stage 0 proposal of audio archive systems in IEC TC 100/TA7 (Teikyo Heisei U.)	Principles for Long-term Preservation of Digital Records (InterPARES)	Curation Practices for the Digital Object Lifecycle, Part II: Addressing Professional Competency Needs through the DigCCurr Professional Institutes (UNC)
Background Which Standard Body			Participated actively in ISO/IEC WG11 (MPEG) and WG1 (JPEG). Main development and contribution was the standardization of the MPEG Query Format which aims on supporting interoperable search requests to multimedia databases and the JPSearch project of JPEG which standardizes interfaces, and protocols for image search.			N/A	
Problem Areas		•	Overcome the diversity in heterogeneous multimedia repositories and provide a standardized access to them. • Metadata interoperability			Record keeping • Authenticity • Reliability • Accuracy	

Audience			<ul style="list-style-type: none"> • Query Language interoperability 				
	<ul style="list-style-type: none"> • 		All multimedia preservation content provider willing to share and cooperate with others.			Organizations, individuals who rely on records to protect rights, obligations, preserve cultural and scientific achievements	
Standard Novelty	<ul style="list-style-type: none"> • 		To provide precise users/systems' input and output parameters in order to express multimedia requests and to allow clients a standard interpretation for processing with the searched result sets. Moreover, MPQF management also provides searching and selection of desired multimedia services to describe service capabilities and to undertake service discovery.			<ul style="list-style-type: none"> • Cross-domain relevance (Govt, scientific and cultural communities; • Multi-disciplinary research (organizational, cultural, scientific) • Historical derivation (archives, diplomatic) • International research 	
Architecture or workflow	<ul style="list-style-type: none"> • 		Standard client-server messaging model for precise input searching parameters and desired output searched results.			Research designed around record keeping lifecycle	
Areas of Application	<ul style="list-style-type: none"> • • 	<ul style="list-style-type: none"> • 	Distributed heterogeneous multimedia search			Long-term preservation of authentic digital records	

Community	•		All organizations handling with multimedia data			<ul style="list-style-type: none"> • Public administration • private archives • science • Culture • Audit • Legal • law enforcement • Information Technology 	
Impact Benefits	•	•	Provide precise searching and unify accessing to heterogeneous multimedia repositories			Governance, law, art, science and scholarship urgently require concrete plans for the preservation of digital materials, so that today's actions, thoughts, achievements and creations will have a future and the future will have a memory.	
Operation Enhancement	<ul style="list-style-type: none"> • • 	<ul style="list-style-type: none"> • • 	<ul style="list-style-type: none"> • Clients: better search results • Repositories: provide standard access to the repository content 			<ul style="list-style-type: none"> • General theory and methods become effective practice; • Specific implementations appropriate to the records in each context; • Skills professionals will require conducting such operations. 	
Cost Saving	•		<ul style="list-style-type: none"> • Eliminate support of multiple priperority accessing interfaces 			N/A	

			<ul style="list-style-type: none"> • Reduce searching and processing time 				
Development Lessons Learned	<ul style="list-style-type: none"> • • 	<ul style="list-style-type: none"> • 	<ul style="list-style-type: none"> • Some query types need to be refined. • Process of result aggregation need to be clarified precisely 			N/A [InterPARES is a research organization, not a standards making body]	
Future Plan	<ul style="list-style-type: none"> • • 		Starting to develop a multimedia middleware supporting the access to multiple heterogeneous repositories			N/A	