Jurnal Internasional Kurasi Digital

Edisi 2, Volume 2 | 2007

Teori dan Aplikasi Pelestarian Digital: Kegiatan Uji Arsip Persisten Transcontinental

Paul Watry,

Universitas Liverpool

November 2007

Abstrak

Proyek Administrasi Arsip dan Catatan Nasional (NARA) dan UE SHAMAN bekerja dengan beberapa lembaga penelitian tentang alat dan teknologi yang akan menyediakan sarana yang komprehensif, sistematis, dan dinamis untuk melestarikan hampir semua jenis catatan elektronik, bebas dari ketergantungan pada perangkat keras tertentu. atau perangkat lunak. Makalah ini menjelaskan kerja pengembangan bersama antara University of Liverpool dan San Diego Supercomputer Center (SDSC) di University of California, San Diego pada prototipe NARA dan SHAMAN. Tujuannya adalah untuk menyediakan teknologi yang mendukung infrastruktur manajemen data umum yang diperlukan. Kami menjelaskan Teori Pelestarian yang mengukur bagaimana komunikasi dapat dilakukan ketika teknologi masa depan berbeda dari yang tersedia saat ini. Ini tidak hanya mencakup perangkat keras dan perangkat lunak yang berbeda, tetapi juga standar yang berbeda untuk informasi pengkodean. Kami menjelaskan konsep "ontologi digital" untuk mencirikan proses pelestarian; ini adalah kemajuan dari Model Referensi OAIS saat ini dalam memberikan informasi representasi tentang catatan. Untuk mewujudkan Teori Pelestarian yang komprehensif, kami menjelaskan integrasi berkelanjutan dari teknologi manajemen koleksi bersama terdistribusi, penelusuran perpustakaan digital, dan teknologi presentasi untuk NARA dan SHAMAN Persistent Archive Testbeds.

Itu *Jurnal Internasional Kurasi Digital* adalah jurnal internasional yang berkomitmen untuk keunggulan ilmiah dan didedikasikan untuk kemajuan kurasi digital di berbagai sektor. ISSN: 1746-8256 IJDC diterbitkan oleh UKOLN di University of Bath dan merupakan publikasi dari Digital Curation Center.

pengantar

Teori Pelestarian

Tuntutan teori pelestarian didorong oleh kebutuhan untuk mengembangkan teknologi pengelolaan data dan informasi yang dapat digunakan untuk membangun arsip yang persisten. Selama bertahun-tahun, jaringan data, perpustakaan digital, dan komunitas arsip persisten masing-masing berfokus pada aspek individual dari masalah, terutama pada pendekatan berbasis metadata yang ditetapkan dalam Model Referensi Open Archival Information System (OAIS). [1]

Standar OAIS berfokus pada kemampuan untuk mengakses dan menafsirkan catatan melalui pembuatan informasi. Namun, tidak memberikan informasi representasi tentang pelestarian lingkungan. Pekerjaan baru-baru ini dimulai pada pengembangan pendekatan berbasis aturan yang memberikan karakterisasi proses pelestarian yang lebih lengkap (Moore & Smith, 2007). Pendekatan ini, dapat dikatakan, akan memberikan cara baru untuk memigrasi semua proses pelestarian (bukan hanya metadata) ke teknologi baru. Hasilnya, arsiparis akan dapat berinteraksi dengan teknologi dan sistem yang tidak diketahui di masa depan sehingga informasi yang berpotensi dapat ditafsirkan dan ditampilkan, menjamin keaslian dan integritas, dari waktu ke waktu.

Sebuah teori pelestarian memperluas konsep pelestarian digital dari yang difokuskan pada pengiriman catatan (metadata) ke masa depan ke yang juga dapat mengirimkan ke masa depan deskripsi lingkungan yang digunakan untuk mengelola dan membaca catatan. Ujian sebenarnya dari lingkungan pengawetan adalah apakah ia menjelaskan seluruh konteks informasi pengawetan dengan cukup baik sehingga rekaman dapat dipindahkan ke lingkungan pengawetan independen tanpa kehilangan keaslian atau integritas. Ini membutuhkan migrasi tidak hanya catatan, tetapi juga karakterisasi konteks lingkungan pelestarian. Lingkungan pengawetan baru harus menerapkan kebijakan pengelolaan yang sama, proses pengawetan yang sama, menggunakan ruang nama logis yang sama, dan mengelola informasi status persisten yang sama. 2007).

Audit & Sertifikasi Repositori Tepercaya RLG / NARA (TRAC): Kriteria dan Daftar Periksa adalah salah satu pernyataan paling canggih dari teori pelestarian. Kriteria memisahkan metadata pengawetan menjadi atribut pada sumber daya penyimpanan, pengguna, koleksi dan data, dan memerlukan kemampuan untuk memelihara konteks informasi, pengaturan, dan deskripsi dari manajemen catatan yang komprehensif [2]. Akibatnya, ini memperluas pendekatan berbasis metadata OAIS menjadi salah satu yang juga dapat mendukung informasi representasi untuk lingkungan pelestarian. Ini membutuhkan mekanisme abstraksi untuk mempertahankan sifat pengawetan meskipun terjadi perubahan yang disebabkan oleh evolusi teknologi. Mekanisme abstraksi yang diperlukan sekarang sedang dikembangkan oleh komunitas grid dan mencakup karakterisasi struktur digital dan semantik; karakterisasi operasi standar pada repositori penyimpanan; karakterisasi kebijakan manajemen; dan karakterisasi mekanisme akses standar.

[1] http://nost.gsfc.nasa.gov/isoas/

[2] http://www.crl.edu/PDF/trac.pdf

Dalam ekspresi persyaratan pelestarian, Kriteria TRAC sangat berpengaruh dalam menetapkan agenda untuk sistem dan proyek pelestarian terkait. Pemetaan termasuk kriteria katalog TRAC / NESTOR untuk penyeberangan repositori digital tepercaya [3], Metode Audit Repositori Digital Berdasarkan Penilaian Risiko (DRAMBORA) [4], dan sistem manajemen data berbasis aturan iRODS [5].

Kriteria TRAC juga menginformasikan kebijakan Digital Preservation Europe (DPE) di bidang konteks pelestarian lingkungan (Hedstrom, 1991).

Makalah ini mengulas konsep otomatisasi proses pengarsipan, dengan fokus pada generasi baru sistem manajemen koleksi berbasis aturan untuk mencirikan struktur catatan. Kami membahas bagaimana sistem ini dapat digunakan untuk mendukung teori pelestarian melalui penerapan "ontologi digital", yang dapat digunakan untuk merepresentasikan hubungan struktural, semantik, spasial, dan temporal yang melekat dalam rekaman (misalnya konteks relatif terhadap produksinya). Kami membahas bagaimana pengerjaan ontologi digital dimajukan melalui pengembangan teknologi objek digital sebagai bahasa deskripsi (“DFDL”) dan alat presentasi (“Multivalent”) yang mengaplikasikan ontologi digital untuk menginterpretasikan rekaman. Kami menghubungkan perkembangan ini dengan upaya penelitian yang sekarang dilakukan untuk mendukung kearsipan berbasis pengetahuan.

Makalah ini dibagi menjadi tiga bagian. Bagian pertama akan membahas agenda penelitian fundamental yang diperlukan untuk mengelola evolusi teknologi. Bagian kedua akan melakukan penilaian terhadap teknologi manajemen informasi yang tersedia saat ini yang dapat digunakan untuk merealisasikan teori pelestarian. Bagian ketiga akan membahas integrasi teknologi untuk mencapai tujuan ini dalam prototipe pelestarian NARA dan SHAMAN [6].

Kegiatan Penelitian Fundamental

Agenda penelitian saat ini difokuskan pada pendefinisian serangkaian proses pengawetan dan atribut pengawetan yang harus dikelola oleh lingkungan pengawetan. Kegiatan dapat dibagi menjadi tiga bidang besar:

- Yang pertama ("data") berfokus pada penggunaan teknologi data grid untuk memetakan informasi representasi OAIS ke ruang nama logis. Ini akan memberikan virtualisasi data dan kepercayaan yang diperlukan untuk kemandirian infrastruktur. Saat ini kami dapat melakukan ini dengan menggunakan teknologi yang ada, misalnya grid data Storage Resource Broker (SRB). Yang kedua ("informasi") berfokus pada karakterisasi proses pelestarian sebagai "ontologi digital" (atau
- informasi representasi) yang mengatur hubungan yang diperlukan untuk menafsirkan struktur dan makna entitas digital. Ini memberikan kemampuan untuk menerapkan label semantik ke struktur dan mengidentifikasi komunitas pengetahuan yang memahami label.

- Yang ketiga ("pengetahuan") berfokus pada karakterisasi manajemen

[3] http://www.digitalpreservationeurope.eu/resources/?search%5B%5D=42

[4] http://www.repositoryaudit.eu

[5] Reagan Moore telah mengembangkan karakterisasi kriteria penilaian TRAC sebagai aturan yang dapat diterapkan oleh grid data iRODS.

[6] SHAMAN: "Mempertahankan Akses Warisan Melalui Pengarsipan Multivalen". Proyek UE terkait termasuk CASPAR: “Pengetahuan Budaya, Seni, dan Ilmiah untuk Pelestarian, Akses, dan Pengambilan” http://www.casparpreserves.eu ; PLANET "Pelestarian dan Akses Jangka Panjang melalui Layanan Jaringan" http://www.planets-project.eu

kebijakan dalam hal aturan dan proses pelestarian sebagai layanan mikro standar. Ini akan memberikan kemampuan untuk menggambarkan evolusi lingkungan pelestarian, dan baik hubungan prosedural yang mengontrol penerapan layanan mikro, dan hubungan fungsional yang membentuk setiap layanan mikro. Ini mengantisipasi kemampuan untuk mengkarakterisasi konten informasi dan pengetahuan untuk presentasi oleh aplikasi baru. Pendekatan ini dirancang untuk mendukung tingkat abstraksi untuk data, informasi, dan manajemen pengetahuan dalam arsip persisten.

Karakterisasi Data

Pendekatan terkini untuk pelestarian digital berfokus pada kemampuan untuk mengakses dan menafsirkan catatan melalui pembuatan "informasi representasi" (yaitu, informasi yang diperlukan untuk merender, menafsirkan, dan memahami data digital) seperti yang didefinisikan dalam Sistem Informasi Arsip Terbuka (OAIS) Model Referensi (ISO 14721: 2003) [7].

Informasi representasi OAIS mendefinisikan struktur dan label semantik dari struktur yang ada di dalam catatan dan komunitas OAIS mempertahankan kemampuan untuk menafsirkan label struktural dan semantik [8]. Paket Informasi Arsip (AIP) OAIS mendefinisikan informasi representasi, Paket Informasi Pengiriman (SIP) mendefinisikan jaminan kualitas, dan Paket Informasi Diseminasi (DIP) mendefinisikan penemuan informasi.

Komunitas jaringan data telah merekayasa perangkat lunak untuk mendukung pendekatan OAIS. Ruang nama logis menyediakan abstraksi yang diperlukan untuk mengelola metadata dalam lingkungan kisi federasi. Ini termasuk karakterisasi struktur digital dan semantik; karakterisasi operasi standar pada repositori penyimpanan; dan karakterisasi mekanisme akses standar. Kisi data Storage Resource Broker (SRB) adalah contoh perangkat lunak yang menyediakan abstraksi untuk kumpulan data, koleksi, pengguna, sumber daya, dan metode proxy yang diperlukan untuk metadata OAIS. SRB mendukung virtualisasi data, atau kemampuan untuk mengelola properti dari koleksi bersama secara independen dari sistem penyimpanan tempat data berada. Pendekatan berbasis gridnya memberikan sejumlah konsep penting untuk manajemen data terdistribusi, termasuk replikasi data (atau akses seragam ke berbagai sumber daya penyimpanan terdistribusi yang heterogen termasuk sistem manajemen basis data, sistem penyimpanan arsip, dan toko file). Berbagai jenis sumber daya penyimpanan dapat digabungkan menjadi lingkungan pelestarian. Ini adalah persyaratan yang jelas untuk mengakses, memelihara, dan berbagi metadata di lingkungan pelestarian.

Pengembangan SRB dan penggandengannya ke model referensi OAIS terutama direkayasa untuk mendukung iterasi pertama Prototipe Arsip Persisten Transkontinental Administrasi Arsip dan Catatan Nasional (NARA) di Washington

DC, dan telah menginformasikan pengembangan infrastruktur pelestarian lainnya. Pengalaman dalam menggunakan perangkat lunak SRB, bagaimanapun, menunjukkan bahwa sebagian besar inisiatif pelestarian digital - termasuk Strategi Penerapan Metadata Pelestarian OCLC (PREMIS) [9], Panduan Persyaratan Data Siklus Hidup NARA (LCDRG) [10].

[7] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683
[8] http://www.dcc.ac.uk/events/jorum-2006/JORUM_oais-08022006.ppt
[9] http://www.oclc.org/research/pmwg
[10] http://www.archives.gov/research/arc/data.html

Proyek Pelestarian Digital SHERPA [11], Portico [12], dan Model Referensi OAIS (antara lain) - didasarkan pada asumsi bahwa pengelolaan metadata pengawetan cukup untuk memelihara lingkungan pengawetan yang lengkap. Untuk menjamin integritas dan keaslian, kami percaya bahwa lingkungan pelestarian perlu mendefinisikan bagaimana proses pengawetan yang diterapkan saat ini dan di masa depan terkait dengan proses pengawetan yang diterapkan di masa lalu. Mengandalkan hanya pada pengelolaan metadata - seperti yang didefinisikan dalam Model Referensi OAIS -, dalam pandangan kami, tidak cukup untuk membuat pernyataan tentang keterpercayaan. Secara efektif, kita perlu mengirimkan ke masa depan tidak hanya informasi (catatan), tetapi juga deskripsi lingkungan ("konteks") yang digunakan untuk mengelola dan membaca catatan (Moore, Arcot, & Marciano, 2007 , hal. 5).

Sebaliknya, Kriteria Audit dan Sertifikasi Repositori yang Dapat Dipercaya (TRAC) telah menegaskan definisi yang lebih lengkap tentang lingkungan pengawetan yang mencakup proses pengawetan dan kebijakan manajemen catatan. Penyediaan informasi ini akan memungkinkan kami untuk mengukur operasi yang diizinkan, mempertahankan operasi yang diizinkan secara independen dari pilihan infrastruktur pelestarian, dan melacak penerapan operasi yang diizinkan. Mekanisme sekarang diperlukan untuk mendeskripsikan metadata OAIS serta lingkungan yang digunakan untuk mengelola dan membaca catatan. Ini akan membutuhkan definisi abstraksi untuk mengkarakterisasi sifat sistemik tentang lingkungan pelestarian.

Sistem Data Berorientasi Aturan (iRODS) terintegrasi yang sedang dikembangkan di San Diego Supercomputer Center (SDSC) dirancang untuk mendukung virtualisasi kebijakan manajemen saat ini, kemampuan pelestarian, dan informasi status yang persisten sambil mempertahankan kemampuan untuk menjalankan kebijakan manajemen sebelumnya. Otomatisasi kebijakan manajemen akan memungkinkan untuk menjadwalkan dan melaksanakan proses yang mendukung penemuan informasi dan manajemen pengetahuan dalam lingkungan pelestarian. Dengan dukungan jaringan data yang sesuai, kita dapat beralih dari situasi berbasis OAIS saat ini - yang mempertimbangkan preservasi dari perspektif format data standar dan mekanisme metadata sederhana - ke model konseptual tingkat tinggi yang mendukung karakterisasi fungsi yang diimplementasikan oleh proses preservasi dan aturan prosedural yang mengontrol penerapan proses. Ini disebut sebagai "arsip berbasis pengetahuan" (Ludascher, Marciano, & Moore, 2001a).

Secara kolektif, tujuan dicapai melalui integrasi sistem manajemen pengumpulan berbasis aturan (iRODS) - dibahas di atas - dengan teknologi Mesin Virtual (Multivalent) yang dapat digunakan untuk menyajikan dan memanipulasi objek dari bitstream asli, tanpa perlu memigrasi atau meniru data. Kami kemudian dapat menjalankan layanan penemuan atau manajemen pengetahuan di lingkungan pelestarian sementara, pada saat yang sama, memastikan keaslian dan integritas data. Konvergensi teknologi dan standar harus mengarah pada dukungan yang lebih baik untuk model struktural dan semantik dari koleksi yang diarsipkan. Ini termasuk representasi relasional "diratakan" ("data"), representasi terstruktur ("informasi"), dan tingkat yang lebih tinggi

[11] http://www.sherpadp.org.uk/index.html
[12] http://www.portico.org/

representasi semantik ("pengetahuan"). Tujuannya adalah untuk mempertahankan kemampuan untuk menemukan, mengakses, dan menganalisis objek digital saat sistem perangkat lunak pendukung berkembang.

Bagian berikut membahas secara lebih mendetail tentang konsep informasi dan konten pengetahuan yang diterapkan pada data, seperti struktur, semantik, konteks, asal, dan properti tampilan. Kami berpendapat bahwa tidak cukup hanya menyalin data pada tingkat bit dari media lama ke media saat ini, tetapi pengarsip juga harus membuat representasi arsip yang dapat dipulihkan yang tidak tergantung infrastruktur (Ludascher, Marciano, & Moore, 2001b , hlm. 9-16). Dengan kata lain, alih-alih memigrasi objek digital melalui pengelolaan metadata OAIS, kami malah mengelola evolusi teknologi. Hal ini menjamin bahwa media penyimpanan, sistem penyimpanan, backup database, dan format objek digital tidak akan menjadi usang, melainkan dapat digunakan untuk menghasilkan pengetahuan baru.

Arsip Persisten Berbasis Pengetahuan

Tantangan utama bagi arsiparis adalah melestarikan makna untuk generasi mendatang. Upaya untuk memenuhi tantangan ini memperkenalkan pertimbangan filosofis dan epistemologis tentang bagaimana merepresentasikan apa pun dalam batas-batas yang dapat diungkapkan oleh bahasa. Sudah diakui bahwa tujuan yang diminta oleh beberapa peneliti pelestarian - kelengkapan koleksi - tidak layak karena kelengkapan adalah penilaian nilai yang tidak dapat diungkapkan secara objektif (Gladney, 2002 , catatan 11). Tidak ada sistem yang dapat mendeskripsikan dirinya sendiri sejauh ini. Sebagai gantinya, upaya penelitian perlu menentukan sekumpulan asumsi minimal yang dapat digunakan untuk mengekspresikan lingkungan pelestarian, sehingga akan mendukung representasi informasi dan pengetahuan sebagai bagian integral dari arsip dan proses penyerapan / migrasi (dikenal sebagai "self- instantiating arsip berbasis pengetahuan ") (Ludascher, Marciano, & Moore, 2001b , hlm. 9-16).

Sejauh mana hal ini dimungkinkan telah menimbulkan banyak perdebatan dalam komunitas pelestarian digital. Pandangan umum yang dipegang adalah bahwa pengelolaan informasi semantik terstruktur dan tingkat yang lebih tinggi merupakan komponen kunci dari penelitian pelestarian digital. Pengarsip memiliki tantangan untuk melestarikan makna semantik dari istilah yang mereka gunakan dalam koleksi untuk mendukung penemuan catatan individu (Gladney, 2002 , catatan 9). Ini akan membutuhkan kemajuan dalam memanipulasi informasi terstruktur dan mengkarakterisasi kebijakan manajemen data yang dapat dibangun di atas gabungan iRODS dan pendekatan Multivalent. Pertanyaan kunci untuk penelitian adalah apakah karakterisasi informasi dan pengetahuan yang lengkap dalam lingkungan pelestarian dapat dicapai, dan apakah ini akan memenuhi kebutuhan arsiparis.

Penelitian sekarang sedang berlangsung untuk menentukan apa yang dimaksud dengan lingkungan pelestarian dan bagaimana kaitannya dengan dunia luar. Lingkungan pengawetan yang benar akan membutuhkan properti pengawetan untuk dipelihara secara independen dari perubahan yang terjadi di dunia luar. Sejauh mana hal ini mungkin berkaitan dengan proses penyerapan ke dalam lingkungan pelestarian dan jaminan bahwa informasi yang memadai dipertahankan untuk memastikan kemandirian infrastruktur lengkap untuk proses pelestarian. Ini adalah pertanyaan yang berkaitan dengan filsafat matematika dan ilmu kognitif. Pekerjaan terbaru di bidang ini berfokus pada bidang web semantik, kecerdasan buatan, dan sistem berbasis aturan untuk mengontrol konsumsi catatan ke dalam lingkungan pelestarian.

Karakterisasi Informasi dan Konten Pengetahuan

Tantangan dalam mengelola entitas digital bukan hanya pengelolaan bit data, tetapi juga pengelolaan infrastruktur yang diperlukan untuk menafsirkan, memanipulasi, dan menampilkan entitas atau gambar realitas ini.

Kita dapat menggunakan spesifikasi berbasis ilmu komputer untuk menggambarkan apa yang diwakili oleh data, informasi, dan pengetahuan. Dalam istilah yang paling sederhana: data sesuai dengan bit (nol dan satu) yang membentuk entitas digital; informasi sesuai dengan setiap tag yang terkait dengan bit. Tag diperlakukan sebagai atribut yang memberikan arti semantik pada bit; pengetahuan sesuai dengan setiap hubungan yang didefinisikan antara atribut informasi. Jenis hubungan terkait erat dengan model data yang digunakan untuk mendefinisikan entitas digital.

Setidaknya, hubungan ini mungkin logis (istilah semantik yang dapat dipetakan ke dalam ontologi dan penalaran dilakukan pada atribut yang disimpulkan); temporal (struktur mungkin mewakili cap waktu yang dapat digunakan untuk menerapkan hubungan kausal); spasial (struktur mungkin mewakili sistem koordinat yang dapat dipetakan ke geometri dan ditampilkan dalam sistem GIS); prosedural (struktur mungkin mewakili hasil dari suatu proses dalam alur kerja); fungsional (struktur mungkin mewakili hasil penerapan algoritma transformasi); atau epistemologis (struktur mungkin mewakili properti sistemik dari seluruh lingkungan pelestarian) (Moore, Arcot, & Marciano, 2007 , hlm. 6). Hubungan semacam itu menentukan konteks informasi lingkungan pelestarian yang dapat dibawa ke depan untuk menghasilkan pengetahuan dari arsip (Boisvert & Tang, 2001).

Teori pelestarian dapat mengkarakterisasi informasi dan konten pengetahuan dalam istilah "ontologi digital" yang dapat digunakan untuk mendefinisikan atribut dan memberikan makna semantik pada data. Atribut dapat ditandai sebagai bagian dari objek digital atau dikaitkan dengan objek digital. Rekaman dan konteks pemrosesannya dipertahankan dan dapat dipindahkan ke teknologi baru. Ini akan memungkinkan untuk menerapkan kembali proses pengarsipan, menjamin bahwa tidak hanya hasil dari proses pengarsipan yang dapat dipertahankan, tetapi juga bahwa deskripsi penerapan proses pengarsipan dapat dipertahankan.

Konvensi penamaan yang digunakan untuk menetapkan arti semantik saat ini sedang ditentukan. Upaya penelitian NARA telah mengembangkan middleware jaringan data iRODS untuk mengkarakterisasi proses pengawetan yang diterapkan pada catatan. Upaya penelitian SHAMAN akan menggunakan Data Format Description Language (DFDL) untuk mendefinisikan dan memberi nama struktur yang ada dalam record dan akan menggunakan model objek digital Multivalent. [13]

untuk mengurai karakterisasi DFDL dari struktur. Tujuannya adalah untuk menggunakan perangkat lunak iRODS untuk mengelola repositori informasi dan perangkat lunak Multivalent untuk memberikan representasi logis dan fisik dari entitas digital, berdasarkan ontologi digitalnya. [14]

[13] http://multivalent.sourceforge.net

[14] Konsep manajemen berbasis aturan dan ontologi terpisah, meskipun saling terkait. Dengan demikian, ontologi dapat mendeskripsikan struktur dan hubungan yang ada dalam rekaman (ontologi ini merupakan perpanjangan dari informasi representasi OAIS). Aturan yang mengontrol penerapan layanan mikro yang menerapkan proses preservasi juga dapat dianggap sebagai hubungan prosedural: himpunan hubungan prosedural dapat diatur sebagai ontologi, tetapi juga dipandang sebagai himpunan aturan. Layanan mikro dapat dilihat sebagai hubungan fungsional yang diberlakukan selama proses pelestarian: sekali lagi, hubungan fungsional dapat diatur sebagai ontologi. Dengan demikian, kita bisa memetakan antar implementasinya

Metode untuk mengkarakterisasi data ilmiah masih kurang maju dan diperlukan pengembangan lebih lanjut dari bahasa deskripsi format data untuk mengurai format data ilmiah (keluaran biner dari kode aplikasi). Selain itu, kami juga memerlukan penambahan tag ke struktur yang menunjukkan unit fisik standar (hubungan fungsional), sistem koordinat (hubungan struktural), geometri (hubungan spasial), dan stempel waktu (hubungan temporal). Pekerjaan di bidang ini sedang berlangsung di komunitas rahasia untuk pelabelan struktur data ilmiah (misalnya dukungan HDF5 untuk kelompok dan array multidimensi), dan beberapa pekerjaan pendahuluan sekarang mulai mencirikan hubungan antara struktur (misalnya model data Konsorsium Open GIS untuk mencirikan transformasi koordinat).

Sifat dan skala tantangan tercermin dalam Networking and Information Technology Research and Development (NITRD) suplemen untuk Anggaran Presiden AS untuk FY2007 [15]. Ini menyoroti pemeliharaan dan akses ke koleksi data sains dan teknik berumur panjang serta catatan Federal sebagai prioritas penelitian. Solusi untuk tantangan tersebut hanya ada sebagian. Diperlukan penelitian di seluruh komunitas terdistribusi yang mengoordinasikan upaya mereka untuk menyelesaikan masalah.

Penilaian Teknologi

Tujuan membangun koleksi bersama dalam lingkungan pelestarian adalah fokus utama dari prototipe pelestarian NARA dan SHAMAN. Ini akan membutuhkan infrastruktur generik yang akan mendukung proses preservasi dan pengelolaan data di beberapa repositori lokal yang didistribusikan di berbagai institusi. Keluaran bersama dari prototipe pengawetan NARA dan SHAMAN akan menghasilkan federasi internasional dari koleksi bersama yang mendukung representasi terstruktur dan semantik dari data.

Prototipe pelestarian NARA dan SHAMAN bertujuan untuk memecahkan tantangan ini melalui penggunaan jaringan data yang menyediakan serangkaian layanan virtualisasi untuk memungkinkan pengelolaan data yang didistribusikan di beberapa situs dan sistem penyimpanan dan teknologi tampilan (Multivalent) memasok kemampuan untuk menyajikan (melihat) dan memanipulasi informasi terstruktur secara independen dari ketergantungan infrastruktur. Bersama-sama, kedua komponen ini menjadi dasar untuk mengelola koleksi arsip di tingkat data, informasi, dan pengetahuan.

Pendekatan ini merupakan kemajuan yang signifikan pada pendekatan berbasis metadata OAIS, sejauh dapat digunakan untuk mengukur bagaimana komunikasi dapat dilakukan di berbagai lingkungan perangkat keras, lingkungan perangkat lunak yang berbeda, dan standar yang berbeda untuk informasi pengkodean. Kami berpotensi mengirim ke masa depan baik informasi OAIS dan deskripsi lingkungan yang digunakan untuk mengelola dan membaca catatan. Kemampuan untuk menafsirkan dan menampilkan catatan secara independen dari kendala infrastruktur membentuk dasar untuk lingkungan pelestarian untuk mendukung kemampuan manajemen pengetahuan.

(aturan, layanan mikro, dan informasi negara) dan deskripsi implementasi yang mengatur hubungan antara aturan, antara layanan mikro, dan antara informasi negara. Informasi dari Reagan Moore (2007).

[15] http://nitrd.gov/

Untuk memahami evolusi sistem hingga titik ini, dan perkembangannya yang diharapkan, akan memerlukan penilaian sistem informasi, karena terwakili dalam grid data, perpustakaan digital, dan komunitas arsip yang persisten. Ada tujuan umum dari infrastruktur manajemen data umum yang memerlukan integrasi analisis data dan analisis pengetahuan; penelusuran dan presentasi perpustakaan digital; dan manajemen koleksi bersama terdistribusi. Sebagian besar upaya integrasi ini dijelaskan dalam laporan yang dikeluarkan oleh San Diego Supercomputer Center, khususnya tulisan Reagan Moore, dan SDSC telah bekerja dengan berbagai komunitas untuk merekayasa integrasi teknologi ini [16].

Inisiatif utama meliputi:

- Grup penelitian Forum Grid Global untuk memanipulasi data dalam repositori penyimpanan terdistribusi.
- Komunitas Perpustakaan Digital untuk mendukung penemuan, akses, dan analisis materi.
- Komunitas Persistent Archive untuk menjaga kemampuan menampilkan dan memanipulasi objek arsip, sementara teknologi yang mendasari berkembang.

Fokusnya adalah pada model data dan informasi yang diperlukan untuk mengelola dan menggabungkan koleksi dan memigrasikannya ke masa depan. Ini melibatkan penggunaan model informasi untuk mendeskripsikan data; kemampuan untuk membedakan konteks yang dibutuhkan untuk kumpulan data, untuk koleksi, dan untuk akses; dan dukungan untuk interoperabilitas antar sistem perangkat keras dan perangkat lunak heterogen (Moore, 1999).

Kami sekarang mencakup aktivitas setiap komunitas dalam menyediakan alat untuk mengelola evolusi teknologi, memperbarui beberapa proyeksi yang dibuat oleh Reagan Moore dalam laporan Teknis SDSC tahun 2003-06 dan menghubungkannya dengan perkembangan NARA dan SHAMAN yang diharapkan. prototipe pengawetan.

Kisi Data

Lingkungan pelestarian didasarkan pada konsep kemandirian infrastruktur yang disediakan oleh jaringan data. Ini dapat diartikan sebagai kemampuan untuk mengelola properti catatan bersama secara independen dari pilihan infrastruktur perangkat keras. Properti tersebut mencakup konvensi penamaan, kontrol akses, informasi administratif, dan tautan ke informasi asal.

Data grid mendukung federasi, kemampuan untuk bertukar koleksi antara data grid yang dikelola secara independen. Ini biasanya dilakukan oleh proyek yang mengelola koleksi bersama secara internasional. Setiap institusi membangun jaringan data lokal dan menegaskan kurasi lokal atas konten. Kisi data independen kemudian dapat disatukan, dengan pengguna, file, dan sumber daya terdaftar silang antara sistem independen. Dengan demikian, pengguna jarak jauh dapat diberikan izin untuk melakukan operasi pengawetan yang dipilih pada data di kisi data terpisah. Pilihan gaya federasi dipilih bersama oleh administrator kisi data. Contohnya termasuk : arsip pusat, di mana jaringan data jarak jauh mendorong data ke sistem pengawetan umum; grid data master di mana file dalam grid data jarak jauh didistribusikan dari grid data master; grid data berantai di mana record direplikasi dari satu grid data ke grid berikutnya di bawah kendali administratif; dan kisi data peer-to-peer di mana tidak ada yang dibagikan dan hanya

[16] Bagian ini mengacu pada pekerjaan dan pengamatan yang dilakukan oleh Reagan Moore di SDSC dan Robert Chadduck di NARA selama iterasi pertama dari prototipe pengawetan NARA (Moore, 2006).

file yang dapat diakses publik dapat diakses dari situs jarak jauh. Dengan demikian informasi dalam beberapa katalog metadata independen dapat disinkronkan di bawah kendali administrator.

Mekanisme abstraksi yang diperlukan untuk evolusi teknologi sedang dikembangkan melalui komite Open Grid Forum (OGF), yang bertemu tiga kali setahun untuk mempromosikan interaksi antara peneliti dan pelaksana jaringan. Komite yang paling relevan adalah Kelompok Riset Lingkungan Pelestarian [17], yang telah mengidentifikasi komponen teknologi jaringan data yang penting untuk pembangunan arsip persisten (rapat Berlin, Maret 2004). Ini termasuk serangkaian proses yang menegaskan keaslian dan integritas entitas digital yang diawetkan sebagai dasar untuk menentukan apa yang merupakan konteks arsip (misalnya atribut administratif, deskriptif, dan keaslian yang terkait dengan setiap entitas digital). Dokumentasi menyimpulkan bahwa masing-masing atribut ini merupakan hasil penerapan suatu proses atau kumpulan hubungan. Penegasan keaslian kemudian sesuai dengan identifikasi hubungan yang telah dipenuhi [18].

Komite terkait yang aktif dalam Forum Jaringan Terbuka termasuk Grup Riset Transportasi Data [19], yang mendefinisikan standar transportasi untuk memastikan interoperabilitas di seluruh repositori penyimpanan; Kelompok Kerja Akses dan Integrasi Data OGF [20], yang mendefinisikan set operasi yang harus dikembangkan untuk berinteraksi dengan teknologi database; Kelompok Kerja Bahasa Deskripsi Format OGF, yang mendefinisikan Bahasa Deskripsi Format Data (DFDL) untuk menjelaskan fitur format data; dan Kelompok Kerja Arsitektur Protokol Grid [21], yang mencirikan kendala konsistensi yang diperlukan untuk merakit grid yang berfungsi dari layanan yang dibedakan dan register informasi status terdistribusi (Moore, 2003). Sejumlah inisiatif sedang menyelidiki penerapan infrastruktur pertukaran data dan metadata berdasarkan middleware OGSA-DAI yang mendukung eksposur sumber daya data, seperti database relasional atau XML, ke grid. [22]

OGF mengevaluasi sejumlah kisi data saat ini untuk menentukan fungsionalitas arsip persisten berdasarkan hasil komunitas pengguna untuk pengelolaan data di berbagai sumber daya penyimpanan terdistribusi dan heterogen (Moore & Marciano, 2007). Persyaratan kisi data untuk arsip persisten mencakup data yang didistribusikan di berbagai situs dan sistem penyimpanan; data dikelola secara independen dari sistem penyimpanan; pengelolaan properti file yang konsisten; pengenal dan kontrol akses yang gigih; dan lingkungan penyimpanan yang dapat diskalakan. Kemampuan ini mendukung banyak skenario penggunaan umum - misalnya, mengelola replikasi data untuk mengurangi risiko kehilangan data - yang merupakan persyaratan mutlak dari arsip persisten.

Masing-masing jenis infrastruktur ini telah diimplementasikan menggunakan jaringan data Storage Resource Broker (SRB) - yang dikembangkan di University of California (San Diego Supercomputer Center) - yang mengimplementasikan ruang nama logis yang digunakan untuk menentukan pengenal global dan persisten yang merupakan lokasi-independen. Layanan pengarsipan membuat informasi status arsip yang dipetakan ke ruang nama logis. Arsip

[17] http://www.gridforum.org/6_DATA/persist.htm
[18] Risalah rapat, tersedia dari http://forge.ogf.org/
[19] http://www.gridforum.org/6_DATA/transport.htm
[20] http://www.gridforum.org/6_DATA/dais.htm
[21] http://www.gridforum.org/5_ARCH/GPA.htm
[22] http://forge.gridforum.org/projects/dais-wg

memetakan atribut asal, administrasi, deskriptif, dan keaslian ke dalam ruang nama logis. Ini mendukung otomatisasi proses pengarsipan.

Operasi yang didukung oleh SRB memberikan representasi operasi yang dapat digabungkan menjadi kemampuan preservasi. Menggunakan SRB memungkinkan pembuatan koleksi bersama yang dapat didistribusikan di berbagai jenis sistem penyimpanan, yang terletak di beberapa domain administratif. Data mungkin didistribusikan, tetapi diindeks oleh katalog metadata terpusat. Operasi kurasi dapat diterapkan dari jarak jauh dari beberapa situs, tetapi hasilnya terdaftar ke dalam katalog metadata umum.

Analisis pendekatan, bagaimanapun, menunjukkan bahwa operasi yang didukung oleh SRB tidak sesuai langsung dengan kemampuan pelestarian yang ditetapkan TRAC. Lingkungan pelestarian membutuhkan kemampuan untuk mengelola representasi informasi yang terkait dengan lingkungan pelestarian itu sendiri. Ini termasuk karakterisasi kebijakan pengelolaan, proses pengawetan, dan pengawetan metadata administratif. Kurator mensyaratkan pengelolaan kedua set properti, dari catatan dan lingkungan pelestarian. Untuk melakukan hal ini diperlukan karakterisasi metadata, proses pengawetan, dan kebijakan manajemen pengawetan terkait. Pendekatan ini diimplementasikan dalam Sistem Data Berorientasi Aturan (iRODS) terintegrasi dari San Diego. [23]

Dalam prototipe arsip persisten NARA dan SHAMAN, iRODS digunakan untuk menunjukkan bahwa pemetaan atribut pelestarian dapat dikelola secara konsisten, melalui aturan yang mengatur struktur dan hubungan yang diperlukan untuk memaksakan pembaruan informasi yang konsisten. Kemampuan untuk mengelola hubungan memungkinkan untuk menerapkan kembali proses pengarsipan, menjamin bahwa tidak hanya hasil dari proses pengarsipan dapat dipertahankan, tetapi juga deskripsi proses pengarsipan

bisa dilestarikan. Penerapan ontologi digital pada entitas digital untuk mengatur hubungan yang diperlukan untuk menafsirkan struktur dan maknanya memungkinkan untuk menjamin kemampuan memanipulasi entitas digital di masa depan. [24].

Munculnya sistem iRODS, yang mengintegrasikan manajemen ontologi dengan manajemen informasi, menunjukkan manajemen hubungan yang didefinisikan secara dinamis antara atribut metadata untuk mendukung federasi namespace. Pemetaan secara konsisten dikelola melalui ontologi yang mengatur hubungan yang diperlukan untuk memaksakan pembaruan informasi yang konsisten. IRODS menyediakan kemampuan untuk mengelola hubungan yang memungkinkan untuk menerapkan kembali proses pengarsipan. Ini menjamin bahwa tidak hanya hasil dari proses kearsipan yang dapat dipertahankan, tetapi juga deskripsi penerapan proses kearsipan dapat dipertahankan. Itu memungkinkan pengembangan ontologi digital sebagai pendekatan migrasi / emulasi gabungan untuk pelestarian (Moore, Arcot, & Marciano, 2007).

[23] R. Moore, dkk, Laporan Teknis SDSC 2003-02r. http://irods.sdsc.edu/index.php/Introduction_to_iRODS

[24]

Teknologi Perpustakaan Digital

Perkembangan dari perpustakaan digital dan komunitas arsip persisten digunakan untuk mendukung pengorganisasian data, akses, dan layanan preservasi, misalnya kemampuan untuk mendukung penemuan dan analisis data bahan yang diorganisir sebagai koleksi. Bidang minat termasuk pengembangan standar metadata untuk objek majemuk; pengembangan standar pertukaran metadata untuk mengambil informasi; perkembangan teknologi untuk pengawetan material; penerapan teknologi pencarian dan manajemen koleksi; dan penerapan teknologi analisis, seperti yang dibahas di bawah ini.

Metadata

Salah satu masalah penelitian utama adalah pengembangan pemahaman tentang berbagai kebijakan pengelolaan yang dibutuhkan oleh masyarakat yang beragam. Prosedur jarak jauh terkait yang dikelola oleh aturan juga bisa unik dan sangat bergantung pada format data. Setiap komunitas akan mengembangkan serangkaian proses dan kebijakan manajemen yang disukai.

Pengalaman saat ini adalah bahwa komunitas pengguna memiliki ekspektasi terhadap properti yang akan dipertahankan oleh koleksi bersama. Mereka biasanya mengharapkan ukuran konsistensi di seluruh catatan (atribut metadata umum), ukuran kelengkapan (tidak ada kumpulan data yang hilang), ukuran keaslian (informasi asal), ukuran integritas (mitigasi risiko terhadap kehilangan data dengan salinan yang valid selalu tersedia). Pada saat yang sama, arsiparis yang mengendalikan lingkungan pelestarian memiliki harapan tentang sifat-sifat yang akan dipertahankan oleh proses pengawetan. Biasanya ada tumpang tindih yang kuat, dengan keinginan untuk keaslian, integritas,

hormati des fonds, lacak balak, mitigasi risiko.

Komunitas pelestarian juga tidak mendamaikan model pelestarian yang sesuai. Setidaknya ada empat model berbeda, termasuk:

- *Diplomatika (InterPARES)* [25]. Komunitas ini berfokus pada keaslian, dengan retensi selamanya;
- *Manajemen data siklus hidup (NARA)* [26]. Komunitas ini berfokus pada jadwal retensi, deskripsi metadata hierarki (grup rekaman, seri rekaman, folder, objek item).
- *Model Kontinum (Universitas Monash)* [27]. Komunitas ini memeriksa pelestarian materi dalam lingkungan yang sama yang digunakan untuk akses dan manipulasi aktif (yang diaktifkan oleh jaringan data). Integrasi perpustakaan digital dan lingkungan pelestarian yang dibahas dalam Pengkajian Teknologi makalah ini akan mengarah pada lingkungan yang mirip dengan model kontinum.

- *Pelestarian Perpustakaan Digital (DSpace)* [28]. Komunitas ini memperluas layanan kurasi menjadi layanan preservasi, namun menggunakan standar perpustakaan digital. Integrasi DSpace dan Fedora ke jaringan data kini telah diperluas untuk memeriksa masalah manajemen kebijakan.

Selain itu, komunitas yang berbeda berfokus pada aspek metadata yang berbeda.

[25] http://www.interpares.org/
[26] http://www.archives.gov/research/arc/lifecycle-data-requirements.doc
[27] http://www.infotech.monash.edu.au/research/groups/rcrg/
[28] http://www.dspace.org/

Untuk komunitas data ilmiah, pengembang alur kerja mencoba mencapai konsensus pada metadata asal yang diperlukan untuk mendeskripsikan pembuatan produk data turunan. Komunitas jaringan semantik malah difokuskan pada pendekatan yang berkembang dari deskripsi layanan berdasarkan kosa kata yang disepakati, menjadi pendekatan yang didasarkan pada karakterisasi fungsi.

Pemetaan dari ekspektasi arsiparis ke ekspektasi komunitas pengguna membutuhkan partisipasi yang kuat dari komunitas pengguna. Komunitas pengguna menentukan format data yang diinginkan, semantik yang diperlukan untuk penemuan, mekanisme akses yang diinginkan, dan kebijakan penggunaan. Lingkungan berbasis aturan membuat karakterisasi kebijakan manajemen ini eksplisit. Aturan yang dibutuhkan oleh setiap komunitas biasanya unik. Kumpulan aturan dapat ditentukan yang diterapkan hanya untuk kumpulan data yang dibuat oleh komunitas itu.

Sebagai langkah awal, berbagai upaya dilakukan untuk memetakan METS dan standar metadata lainnya [29], termasuk direktori data Preservation Metadata Implementation Strategies (PREMIS) [30] dan penyeberangan dengan skema perpustakaan digital, misalnya Dublin Core (ANSI / NISO Z39.85-2007). [31] Hasilnya harus berupa standarisasi metadata deskriptif untuk menentukan keaslian.

Untuk memenuhi persyaratan Trustworthy Repositories Audit and Certification (TRAC), pekerjaan di masa depan mungkin memerlukan METS dan standar metadata lainnya untuk ditambah dengan atribut yang menentukan konteks informasi dari lingkungan preservasi, untuk memungkinkan deskripsi kebijakan manajemen preservasi dimigrasikan Menuju masa depan. Salah satu hasil dari prototipe NARA dan SHAMAN adalah spesifikasi kriteria untuk jenis informasi representasi ini. Ini akan mencakup karakterisasi infrastruktur perangkat keras / perangkat lunak, atau pada karakterisasi tingkat yang lebih tinggi dari operasi yang didukung oleh lingkungan, atau pada karakterisasi tingkat yang lebih tinggi dari layanan mikro, aturan, dan informasi status persisten yang dipertahankan oleh lingkungan pelestarian .

Penelitian lebih lanjut akan diperlukan untuk mengintegrasikan semua aspek ini ke dalam sistem yang koheren. Integrasi perpustakaan digital dan teknologi pelestarian harus menghadapi integrasi layanan kurasi perpustakaan digital (untuk akses dan penemuan) dengan proses pengarsipan pelestarian. Upaya pertama adalah mempromosikan akses sedangkan upaya kedua adalah mempromosikan keaslian. Jadi, dua set metadata yang mungkin terputus-putus mungkin diperlukan.

[29] Metadata Encoding and Transmission Standard (METS) memberikan kemampuan untuk mengkarakterisasi struktur objek digital yang dapat didefinisikan, diatur, dan dipelihara secara independen dari komponen yang terpisah, dengan dukungan untuk metadata asal, metadata administratif, metadata pelestarian, dan metadata struktural : http://www.loc.gov/standards/mets/ Pengarsip semakin tertarik untuk menentukan hubungan antara informasi negara arsip yang terkandung dalam standar METS dan persyaratan standar pelestarian 5015.2 Departemen Pertahanan AS untuk manajemen arsip: http: // http: //jitc.fhu.disa.mil/recmgt/ ; dan standar ISO TR 15801 tentang Manajemen Arsip dan Pertimbangan Hukum http://www.iso.org/iso/catalogue_detail?csnumb

[30] http://www.oclc.org/research/projects/pmwg/

[31] Lihat proyek Eropa DIFFUSE (Penyebaran Spesifikasi dan Pengalaman InFormal dan Formal yang Berguna): http://www.dcc.ac.uk/diffuse/

Dublin Core Metadata Initiative: http://www.dublincore.org/

Alur Kerja Pelestarian

Alur kerja pengawetan digital digunakan untuk mengimplementasikan proses pengawetan dengan catatan yang disimpan dalam sistem manajemen informasi. Prototipe pelestarian NARA menerapkan sistem berbasis aturan (Jaringan Alur Kerja Produser-Arsip
- "PAWN") untuk mengontrol konsumsi catatan ke dalam lingkungan pengawetan (Smorul, JaJa, Wang, & McCall, 2004). Sistem ini dirancang untuk digunakan dengan model referensi OAIS dan grid data untuk merangkum metadata konten, struktural, deskriptif, dan pelestarian.

Baru-baru ini, San Diego telah menggunakan sistem alur kerja Kepler [32] untuk mengintegrasikan proses pengawetan ke dalam siklus produksi produksi konten, menggunakan alur kerja produksi video yang ada (Arcot, Moore, Berman, & Schottlaender, 2005). Proyek ini dirancang untuk mengabstraksi alur kerja produksi dan pengelolaan siklus hidup pelestarian. Sistem alur kerja Kepler juga digunakan untuk mengumpulkan dan memelihara informasi asal untuk data ilmiah dan untuk dasar tantangan asal (Altintas, Barney, & Jaeger-Frank, 2006) [33].

Yang paling menarik adalah perbandingan aturan untuk mengontrol konsumsi dengan aturan yang diperlukan untuk menjaga keaslian dan integritas. Aturan konsumsi harus mengekspresikan interaksi antara lingkungan pelestarian dan dunia luar. Aturan untuk menjaga keaslian dan integritas harus menunjukkan bahwa properti pengawetan dipertahankan secara independen dari dunia luar. Pertanyaannya adalah apakah aturan keaslian dan integritas beroperasi hanya pada informasi status persisten yang dijaga oleh lingkungan pelestarian, dan dengan demikian tidak tergantung pada perubahan yang terjadi di dunia luar. [34] Hasil penelitian dari proyek pelestarian SHAMAN akan menentukan sejauh mana interaksi dengan dunia luar dapat diringkas ke dalam aturan konsumsi, dengan ketergantungan infrastruktur yang dikelola oleh driver yang dibuat untuk mendukung kemandirian infrastruktur.

Teknologi Manajemen Pencarian dan Koleksi

Teknologi perpustakaan digital menyediakan layanan standar untuk penyerapan, akses, dan tampilan berdasarkan standar dan protokol metadata. Integrasi dengan teknologi data grid dapat memastikan bahwa sistem perpustakaan digital memperoleh kemampuan untuk mengelola koleksi yang melebihi ukuran sistem file lokal, mendapatkan dukungan untuk replikasi, dan memperoleh kemampuan untuk menggabungkan dengan perpustakaan digital lainnya.

Proyek integrasi jaringan data perpustakaan digital DSpace-SRB mendemonstrasikan contoh penggabungan manajemen informasi terdistribusi melalui penggunaan perpustakaan digital dan federasi lingkungan pelestarian s melalui penggunaan kisi data. Proyek ini telah menghasilkan dukungan DSpace-SRB untuk sebagian kecil daftar kemampuan Arsip Rekaman Elektronik (ERA), yang dirancang untuk prototipe pelestarian NARA. Integrasi DSpace-SRB dapat digunakan untuk mempertahankan koleksi kecil, tetapi saat ini tidak memiliki kemampuan untuk menskalakan koleksi besar. Penelitian lebih lanjut akan diperlukan untuk menentukan apakah kebijakan manajemen dapat ditentukan di perpustakaan digital DSpace dan didukung oleh jaringan data iRODS.

[32] http://www.kepler-project.org/
[33] http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge
[34] Informasi dari Reagan Moore (2007).

Teknologi perpustakaan digital dapat digunakan untuk mengaitkan fungsi tampilan dengan setiap tipe data, memungkinkan hubungan diterapkan pada rekaman, dan memetakan label semantik pada rekaman ke ontologi digital. Perpustakaan digital Fedora dan integrasi SRB, sebagai bagian dari Dataset Acquisition, Accessibility, and Annotation eResearch Technologies (DART) Project, dirancang untuk mengimplementasikan lingkungan preservasi, dengan fokus khusus pada hubungan logis.[35] Upaya penelitian lebih lanjut diperlukan untuk menentukan bagaimana Fedora dapat digunakan untuk mengelola semantik pada atribut pelestarian.

Standar Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) digunakan untuk mendukung publikasi metadata dari koleksi independen ke dalam repositori pusat.[36] OAI menyediakan mekanisme akses atribut untuk manipulasi oleh program. Teknologi ini memiliki serangkaian layanan pelengkap yang ditentukan dalam Kelompok Kerja Layanan Integrasi dan Akses Database (DAIS) dari OGF[37]. Grup tersebut telah menetapkan bahwa integrasi teknologi perpustakaan digital dan teknologi grid diperlukan untuk membuat standar yang dapat digunakan dalam arsip persisten. Kelompok kerja telah mengusulkan untuk menerapkan antarmuka DAIS yang digunakan untuk berbicara ke database dan antarmuka OAI yang digunakan untuk berbicara dengan registri yang mencantumkan database. Proyek ilmiah yang menyelidiki integrasi ini termasuk National Virtual Observatory (NVO)[38].

Teknologi perpustakaan digital juga dapat digunakan untuk memperkenalkan penemuan tingkat lanjut dan kemampuan analisis data untuk arsip persisten. Menggunakan framework perpustakaan digital Cheshire[39], misalnya, dimungkinkan untuk menerapkan alat analisis pengetahuan dan analisis data ke koleksi bersama yang catatannya berada di berbagai jenis sistem penyimpanan, di berbagai institusi, yang berlokasi di banyak negara (Larson & Sanderson, 2005 , 2006 ; Watry & Larson, 2005). Sistem menyediakan antarmuka ke banyak algoritme penambangan data, termasuk pengelompokan, klasifikasi, dan penambangan aturan asosiasi yang semuanya dapat digunakan untuk koleksi objek dalam lingkungan pelestarian. Penelitian lebih lanjut sekarang diperlukan untuk menyelidiki penggunaan dukungan alur kerja untuk pencarian dan proses manajemen koleksi, menggunakan sistem alur kerja Kepler dan kerangka kerja perpustakaan digital berorientasi layanan Cheshire yang menjalankan alur kerja pemrosesan.

Keluaran dari prototipe pengawetan SHAMAN akan mengevaluasi bagaimana aturan yang diterapkan oleh iRODS dapat digunakan untuk memfasilitasi analisis oleh sistem Cheshire. Ini akan mencakup migrasi otomatis koleksi oleh iRODS ke disk berkinerja tinggi (Teragrid) untuk analisis Cheshire; penerapan layanan Cheshire secara langsung di sistem penyimpanan jarak jauh melalui aturan iRODS; dan penjadwal yang menentukan apakah data harus dipindahkan ke platform komputasi untuk komputasi oleh Cheshire atau apakah layanan harus dijalankan di sistem penyimpanan jarak jauh di bawah aturan iRODS.

Untuk prototipe pengawetan NARA, sistem Cheshire diintegrasikan dengan kisi data SRB untuk mendukung pencarian terdistribusi di lingkungan pengawetan. Sistem Cheshire kini telah ditanamkan dalam hal pemrosesan alur kerja yang bisa

[35] http://ausweb.scu.edu.au/aw06/papers/refereed/treloar/paper.html
[36] http://www.openarchives.org/OAI/openarchivesprotocol.html
[37] http://forge.ogf.org/sf/projects/dais-wg
[38] http://www.adass.org/adass/proceedings/adass03/P3-8/ (di bawah Bagian 3.2).
[39] http://www.cheshire3.org/

mengintegrasikan proses penggalian teks dan data dalam lingkungan pencarian terdistribusi. Untuk prototipe pelestarian SHAMAN UE, sekarang kami akan menyelidiki bagaimana aplikasi komputasi intensif, seperti analisis, dapat dilakukan di sisi klien dengan data yang dimigrasi ke platform yang kuat tempat komputasi dilakukan. Hasil dari ini akan menjadi integrasi alur kerja sisi klien (mendukung pencarian dan teknologi manajemen pengumpulan) dengan alur kerja sisi server dari lingkungan berorientasi aturan untuk mendukung semua skala komputasi dan manipulasi data.

Arsip Persisten (Penyimpanan Data)

Perkembangan teori pelestarian - dari yang difokuskan pada pengiriman metadata ke masa depan ke yang juga dapat mengirimkan ke masa depan deskripsi lingkungan yang digunakan untuk mengelola dan membaca catatan - secara langsung dapat dilacak ke grid dan silsilah perpustakaan digital yang diuraikan di atas. Kemampuan untuk mengkarakterisasi kebijakan pengelolaan dalam hal aturan dan proses pengawetan sebagai layanan mikro standar diperlukan untuk memenuhi strategi pengawetan yang berbeda yang mungkin diperlukan oleh komunitas pengawet yang berbeda.

Lingkungan pengawetan di masa depan harus dapat menentukan strategi pengawetan terpisah untuk setiap rangkaian rekaman (koleksi) berdasarkan integrasi jaringan data dan teknologi perpustakaan digital. Misalnya, jika sebuah koleksi memiliki kebijakan penyimpanan migrasi transformatif ke XML, kita harus dapat menentukan aturan yang menjalankan layanan mikro jarak jauh di sistem penyimpanan tempat rekaman berada yang mengonversi dari format usang ke format baru. Koleksi kedua mungkin menggunakan pendekatan emulasi: dalam hal ini, aturan yang mengontrol tampilan record akan menggunakan versi yang benar dari teknologi emulasi; misalnya, Portable Document Format (PDF). [40] Tujuannya adalah untuk dapat memastikan kebijakan pengawetan mana yang diterapkan untuk setiap koleksi, dan melacak penerapan kebijakan pengawetan dari waktu ke waktu. Ini berarti melacak kumpulan aturan yang diterapkan, kumpulan layanan mikro jarak jauh yang dijalankan oleh kumpulan aturan, dan informasi negara yang dihasilkan dari penerapan aturan. Contohnya adalah jejak audit dari semua migrasi transformatif yang diterapkan.

Untuk mencapai tujuan tersebut, teknologi arsip persisten yang diperlukan untuk prototipe pelestarian NARA dan SHAMAN terutama didasarkan pada aplikasi grid data berbasis aturan iRODS dan teknologi objek digital Multivalent. Yang pertama mendukung pengelolaan hubungan antara atribut metadata; yang terakhir mendukung kemampuan untuk menerapkan hubungan ini untuk menafsirkan entitas digital. Kedua teknologi tersebut dapat digunakan sebagai dasar untuk mendukung kriteria dan checklist Trustworthy Repositories Audit and Certification (TRAC). Sebagai sintesis dari grid data dan teknologi perpustakaan digital yang dibahas di atas, pendekatan iRODS dan Multivalent dapat digunakan untuk mendukung tidak hanya pengelolaan metadata, tetapi juga deskripsi lingkungan yang digunakan untuk mengelola dan membaca catatan.

[40] Perbedaan antara pendekatan Multivalent dan pendekatan berbasis metadata (misalnya emulasi dokumen, konversi / migrasi, format universal, dan komputer universal) dibahas di Phelps dan Watry (2005).

Untuk arsip persisten, tujuannya adalah menggunakan kisi data iRODS untuk mengelola pemetaan secara konsisten melalui ontologi yang mengatur hubungan yang diperlukan untuk memaksakan pembaruan informasi yang konsisten. (Ini adalah langkah di luar SRB, yang dengan sendirinya dapat menunjukkan kemampuan untuk mengelola pemetaan informasi yang sesuai ke ruang nama logis, tetapi tidak memiliki virtualisasi manajemen.) Kemampuan iRODS untuk mengelola hubungan memungkinkan untuk menerapkan kembali proses pengarsipan, menjamin bahwa tidak hanya hasil dari proses pengarsipan yang dapat dipertahankan, tetapi juga deskripsinya. Penggunaan Multivalent kemudian dapat digunakan untuk menafsirkan struktur entitas digital dan menjamin bahwa mereka dapat dimanipulasi di masa mendatang.

Dalam komunitas pelestarian digital, penelitian di bidang alat presentasi sebagian didasarkan pada konsep Komputer Virtual Universal (UVC) [41],
yang mendefinisikan operasi preservasi pada level bit dan dapat, secara teori, mencirikan manipulasi pada record sebagai operasi level bit. Ini membutuhkan program yang canggih untuk menafsirkan bit dengan benar. Sistem UVC dapat dimigrasi ke sistem operasi baru, secara teori memungkinkan penguraian dan manipulasi catatan ke sistem baru. Namun, penerapan UVC memiliki sejumlah kekurangan yang diakui, terutama dalam cara kebijakan manajemen dikodekan ke dalam perangkat lunak. Saat ini mungkin untuk menggunakan UVC untuk mengurai format yang relatif sederhana, tetapi bukan yang rumit. Agar lebih efektif, pendekatan tersebut perlu mengimplementasikan kemampuan untuk memetakan dari tindakan aplikasi ke operasi tingkat bit sebagai operasi tingkat tinggi generik. Ini akan memberikan kemampuan untuk menangani dokumen dan format data yang kompleks,

Sebaliknya, teknologi objek digital Multivalent sudah dapat mendukung operasi tingkat tinggi generik untuk memanipulasi karakterisasi struktur dalam rekaman. Ini akan memungkinkan interpretasi entitas digital untuk pelestarian dan manipulasi sementara teknologi yang mendasarinya berkembang. Pendekatan Multivalent membedakan antara penguraian melalui adaptor media dan manipulasi melalui penerapan perilaku. Setelah dokumen atau objek digital diurai, itu kemudian dapat dimanipulasi oleh perilaku Multivalent standar. Multivalent dapat mem-parsing berbagai format data (misalnya PDF, OpenOffice, HTML) dan, karena Multivalent ditulis dalam Java, ia dapat di-porting ke sistem operasi baru dengan relatif mudah.

Multivalent dapat dianggap sebagai lingkungan emulasi yang ditulis menggunakan bahasa tingkat yang lebih tinggi (Java) yang memisahkan masalah penguraian dari tampilan, dan yang menyediakan pustaka operasi standar yang dapat digunakan untuk menampilkan dan memanipulasi dokumen dan data. Arsitektur Multivalent dirancang untuk menafsirkan entitas digital berdasarkan ontologi digital yang mewakili hubungan struktural, semantik, spasial, dan temporal yang melekat dalam entitas digital (Phelps & Watry, 2005). Dengan cara ini, ia dapat membuat semua rekaman dari bentuk aslinya dan menjamin interpretasi yang benar dari rekaman tersebut di lingkungan pelestarian di masa mendatang. Ini memberikan bentuk kemandirian infrastruktur untuk aplikasi tampilan.

Karakterisasi ontologi digital ini sedang dilakukan oleh kelompok penelitian Data Format Description Language (DFDL) dari Global Grid Forum, yang sedang mengembangkan deskripsi berbasis XML tentang struktur yang ada di dalam struktur. Multivalent digunakan sebagai alat presentasi yang menerapkan hubungan ini, dalam

[41] http://www.nla.gov.au/padi/topics/492.html

urutan yang ditentukan oleh ontologi digital DFDL, untuk menafsirkan entitas digital. Dengan demikian, ini akan mendukung presentasi dan manipulasi objek digital (misalnya dokumen) tanpa ketergantungan. Contohnya adalah penggunaan Multivalent untuk menampilkan dokumen Adobe Acrobat PDF dari bitstream asli (yaitu tanpa memerlukan perangkat lunak Adobe untuk berada di sistem) [42]. Perkembangan teknologi iRODS memungkinkan untuk merepresentasikan hubungan struktural, semantik, spasial, dan temporal yang melekat dalam entitas digital. Model objek Multivalent kemudian dapat menggunakan informasi ini untuk menafsirkan entitas digital dari bitstream asli.

Untuk jenis data tertentu, "adaptor media" dibuat yang mengubah objek menjadi struktur internal antarmuka Multivalent. Adaptor media adalah komponen kode yang menerjemahkan format dokumen konkret ke dalam struktur data runtime. Struktur data primer adalah pohon dokumen, yang mewakili seluruh konten dokumen (sebagai gulungan atau halaman pada satu waktu) termasuk segala sesuatu mulai dari teks dan gambar, hingga skrip, hingga struktur semantik (hierarki dan atribut), hingga tata letak fisik. Elemen aktif (programatik) atau dokumen tertentu atau genre dokumen, seperti hyperlink atau pembukaan dan penutupan garis besar, diimplementasikan oleh perilaku, yaitu kode program dengan akses lengkap ke konten dokumen. Perilaku tertentu yang berlaku untuk dokumen atau genre dicantumkan di hub format XML.

Untuk prototipe pelestarian NARA dan SHAMAN, teknologi Multivalent (program Java) dan adaptor media diarsipkan, bersama dengan apa pun yang diperlukan untuk memigrasi mesin Virtual Java ke masa depan. Emulasi kemudian terdiri dari mendukung operasi asli untuk memanipulasi entitas digital. Migrasi terdiri dari porting Java Virtual Machine ke teknologi baru sesuai kebutuhan. Tidak seperti strategi migrasi saat ini, entitas digital tetap tidak berubah, sekaligus memungkinkan untuk menerapkan operasi baru yang tersedia di versi baru Multivalent.

Kemampuan pendekatan ini untuk menangani data ilmiah dan teknik terstruktur sekarang mulai dinilai. Seperti yang dinyatakan di atas, prototipe pelestarian NARA dan SHAMAN saat ini terlibat dengan tantangan dalam melakukan kurasi pengumpulan data ilmiah dan teknik, dengan tujuan untuk dapat mengarsipkan data dalam konteks yang dapat memastikan kegunaannya di masa mendatang. Kita perlu mendukung skenario penggunaan masa depan yang mungkin menjadi sesuatu yang tidak diantisipasi oleh pembuat repositori atau pengarsip. Ada sejumlah pertanyaan penelitian yang berkaitan dengan proses tersebut: misalnya, apakah mungkin untuk mendefinisikan sekumpulan perilaku yang menampilkan data ilmiah menggunakan Multivalent?

Skenario Implementasi

Tujuan dari pengujian NARA dan SHAMAN adalah untuk mengembangkan dan menerapkan kebijakan dan perangkat lunak yang akan mencapai tujuan ini dengan memungkinkan pemindahan tidak hanya catatan, tetapi juga karakterisasi konteks lingkungan pelestarian itu sendiri. Ini adalah metrik yang menentukan untuk menjelaskan lingkungan pelestarian yang diwajibkan oleh industri dan pemerintah untuk memenuhi kepatuhan sesuai dengan Arahan UE tentang Privasi dan Komunikasi Elektronik (2002/58 / EC) dan Peraturan Federal untuk Prosedur Sipil (FRCP) yang baru di Amerika Serikat.

Implementasi di bawah NARA terdiri dari tujuh grid data independen yang mengelola data yang didistribusikan di seluruh sumber daya penyimpanan di tujuh institusi (salinan utama disimpan di NARA I di Washington DC; salinan yang direplikasi di University of Maryland untuk

[42] http://multivalent.sourceforge.net ; http://bodoni.lib.liv.ac.uk/fab4/

peningkatan akses dan pemulihan bencana; arsip mendalam di SDSC; laboratorium di NARA II di College Park, Maryland; laboratorium di Rocket Center di West Virginia; kolaborator di Universitas Carolina Utara, Chapel Hill; kolaborator di Georgia Tech). Ini sekarang digunakan untuk mendemonstrasikan manajemen evolusi teknologi, pelestarian web crawl, ekstraksi otomatis metadata keaslian, dan pipeline pengiriman arsip-produsen. Implementasi masa depan untuk prototipe SHAMAN akan terdiri dari tiga grid data independen di Inggris Raya, Portugal, dan Jerman dengan rencana untuk terhubung dengan prototipe NARA untuk membentuk testbed arsip persisten lintas benua.

Meskipun prototipe NARA mewakili keadaan seni terkini dalam teknologi pengawetan digital, namun juga menunjukkan perlunya generasi baru teknologi, yang diperoleh melalui kemajuan penelitian, yang akan memenuhi tujuan pelestariannya, termasuk:

- Keaslian, penegasan bahwa metadata deskriptif asal dan metadata integritas tetap terkait erat dengan catatan elektronik di semua proses pengarsipan, dan bahwa metadata asal belum diubah;

- Integritas, pernyataan bahwa catatan elektronik tidak rusak, dan bahwa rantai pengawasan arsip telah diberlakukan dan diaudit; Kemandirian infrastruktur,
- pernyataan bahwa pelestarian lingkungan dapat dipertahankan melintasi evolusi sembarang komponen infrastruktur

SHAMAN dibangun di atas prototipe NARA dalam pengembangan teknologi iRODS, DFDL, dan Multivalent: upaya bersama akan dilakukan untuk memperluas dukungan untuk representasi data terstruktur dan semantik. Ini akan membutuhkan ekstensi kapabilitas iRODS untuk mencirikan struktur dan hubungan dalam rekaman, mengidentifikasi operasi standar yang dapat dilakukan pada hubungan tersebut, dan memetakan dari tindakan yang dijalankan oleh aplikasi tampilan ke operasi standar [43]. Upaya paralel akan dilakukan untuk memperluas teknologi Data Format Description Language (DFDL) dan Multivalent digital object untuk mem-parsing dan membuat data dan hubungannya. Upaya penelitian akan fokus pada konsep ontologi digital sebagai pendekatan migrasi / emulasi baru untuk pelestarian.

Hasil penelitian dan pengembangan NARA I ditunjukkan dalam arsip persisten prototipe penelitian. Demonstrasi saat ini mencakup akses ke catatan elektronik yang direplikasi di tiga grid data yang terdiri dari arsip persisten menggunakan grid data SRB dan sistem Cheshire; aksesi koleksi sampel melalui pipa produser-arsip PAWN yang dikembangkan di University of Maryland; aksesi penjelajahan web; presentasi arsip menggunakan teknologi Multivalent, dan validasi kepemilikan arsip persisten. Untuk testbeds NARA di masa depan dan prototipe SHAMAN, kami berharap dapat memperluas kemampuan analisis data dan analisis pengetahuan untuk mendukung penemuan di seluruh koleksi data independen melalui integrasi ontologi.

Pendekatan jangka panjang berbasis penelitian NARA dan SHAMAN telah memberikan cara yang komprehensif dan dapat dipercaya untuk menangani apa yang bukan hanya target bergerak, tetapi juga yang berkembang pesat baik secara kuantitatif maupun dalam kompleksitas, dan di sepanjang jalur.

[43] Informasi dari Reagan Moore (2007).

yang tidak sepenuhnya dapat diprediksi (Thibodeau, 2001). Contohnya adalah persyaratan terkini untuk melestarikan konten, struktur, dan konteks di berbagai media, bukan hanya dokumen. Munculnya "Podcasting" dalam beberapa tahun terakhir adalah salah satu contoh yang mungkin tidak diantisipasi pada awal proyek, tetapi perlu diakomodasi dalam iterasi berikutnya dari prototipe dan layanan akhir.

Pada akhirnya, sistem NARA dan SHAMAN harus dapat menjaga konten, struktur, dan konteks data apa pun, termasuk alat analisis khusus yang merupakan karakteristik eScience atau cyberinfrastructure. Ini adalah tantangan utama yang memerlukan adopsi yang lebih baik dari format data yang terdefinisi dengan baik dan semantik yang terdefinisi dengan baik; itu juga akan membutuhkan sarana untuk mencatat asalnya sehingga - misalnya - kondisi yang tepat dari eksperimen ilmiah dapat diulangi dengan penuh keyakinan.

Layanan dasar NARA dan SHAMAN diharapkan terdiri dari 854 kapabilitas sebagaimana ditentukan oleh TRAC. Lebih dari 200 kebijakan manajemen yang berbeda dapat ditentukan untuk mengontrol kapabilitas. Upaya penelitian saat ini mendefinisikan setiap kemampuan dalam hal aturan yang mengatur tindakan yang dilakukan pada data dan pernyataan yang dibuat tentang koleksi (Moore, et al., 2006). Dalam mendukung setiap aturan ini, prototipe NARA menggunakan teknologi jaringan data yang terintegrasi dengan perpustakaan digital dan teknologi pelestarian untuk mengelola:

- aksesi catatan elektronik,
- proses pelestarian standar untuk membuat bentuk arsip dari catatan elektronik dan mengekstrak keaslian dan integritas metadata,
- template untuk menentukan informasi terstruktur preservasi yang diperlukan,
- sistem manajemen alur kerja untuk menerapkan templat ke seri rekaman, dan
- sistem kendala untuk menerapkan kebijakan pelestarian.

Teknologi terintegrasi ini terkait erat dengan pengetahuan dan kemampuan analisis data yang disediakan oleh University of Liverpool. Awalnya dirancang secara independen dari prototipe NARA, alat yang berbasis di Liverpool sekarang menafsirkan entitas digital untuk presentasi dan manipulasi dengan cara yang didasarkan pada teknologi jaringan data SRB dan iRODS. Mereka juga membentuk (bersama dengan teknologi perpustakaan digital terkait seperti Fedora, CITRIS, DSpace) area penelitian aktif, terutama mengenai spesifikasi kebijakan manajemen data perpustakaan digital, yang memerlukan modifikasi mekanisme manajemen jaringan data untuk mendukung manipulasi struktur data. informasi.

Dalam konteks prototipe NARA dan SHAMAN:

- Lingkungan berorientasi aturan (iRODS) menyediakan mekanisme abstraksi untuk mengelola evolusi penyimpanan dan repositori informasi;

- Perpustakaan digital dan sistem analisis pengetahuan mempertahankan kemampuan untuk mengelola, mengakses, dan menganalisis data.
- Teknologi pengawetan Multivalen memberikan kemampuan untuk menafsirkan entitas digital untuk presentasi dan manipulasi sementara teknologi yang mendasarinya berkembang.

Teknologi, secara bersama-sama, mewakili visi bersama untuk pelestarian semua komponen arsip persisten dan menggambarkan kelayakan akses jangka panjang dan tampilan entitas digital. Tantangan dalam perspektif

Prototipe NARA dan SHAMAN adalah untuk mengotomatiskan semua aspek penemuan, akses, manajemen, dan manipulasi data. Meskipun saat ini kami dapat mendemonstrasikan otomatisasi proses pengarsipan dalam skala besar pada akses kepemilikan digital NARA yang ada (dan pendaftarannya ke dalam arsip persisten), kami baru sekarang mulai memahami apa yang diperlukan untuk mengembangkan sistem manajemen pengumpulan berbasis kendala dan tantangan lebih lanjut untuk mengembangkan konsep ontologi digital sebagai pendekatan pelestarian.

Menerapkan Kebijakan Manajemen untuk Memvalidasi Trustworthy Repositories Audit and Certification (TRAC)

Pendekatan pelestarian lanjutan tidak hanya perlu menangani tujuan langsung dari pengelolaan dan pelestarian catatan - memastikan keaslian dan integritas - tetapi juga integrasi kemampuan ini ke dalam infrastruktur yang mendukung kemampuan untuk mengekspresikan kebijakan manajemen, misalnya layanan yang meminta informasi status yang persisten untuk divalidasi kriteria penilaian kepercayaan.

Tantangannya ada empat kali lipat:

- Untuk mengimplementasikan infrastruktur manajemen informasi siklus hidup yang akan menjamin kemampuan untuk memelihara konteks informasi, pengaturan dan pengelolaan catatan;

- Untuk mengimplementasikan kemampuan untuk mengurai, menampilkan, dan memanipulasi objek digital secara independen dari batasan infrastruktur apa pun;
- Untuk mengintegrasikan analisis data dan alat penemuan untuk memungkinkan pengguna dan administrator menemukan informasi yang diinginkan dalam lingkungan ini; dan Untuk memungkinkan
- penciptaan pengetahuan di masa depan melalui pembuatan dan penerapan aturan inferensi dalam mesin inferensi yang dapat diskalakan.

Tujuan jangka panjangnya adalah teori "rekayasa" manajemen data. Kami menguraikan kapabilitas yang diperlukan menjadi serangkaian layanan mikro, menguraikan kebijakan manajemen menjadi serangkaian aturan; menguraikan kriteria penilaian menjadi pertanyaan tentang informasi status yang persisten. Dengan adanya tiga ruang ini, kami kemudian dapat menunjukkan bahwa pemetaan antara ketiga ruang tersebut dalam hal operasi aktual adalah lengkap, konsisten, dan tertutup. Penting untuk membuktikan bahwa lingkungan pelestarian tidak menimbulkan ketergantungan apa pun dalam bahan yang diawetkan pada pilihan teknologi. Cara untuk membuktikan kemandirian infrastruktur adalah dengan memigrasikan materi yang diawetkan ke teknologi pengawetan independen (menggunakan pilihan teknologi alternatif), dan kemudian memindahkan materi yang diawetkan kembali ke lingkungan pengawetan asli tanpa kehilangan keaslian atau integritas.

Ini membutuhkan infrastruktur manajemen koleksi yang dapat memigrasi entitas digital yang diawetkan ke sistem perangkat keras baru tanpa mengubah ruang nama yang digunakan untuk mengelola entitas digital; sistem perangkat lunak baru tanpa mempengaruhi protokol yang digunakan untuk mengakses data; protokol akses baru tanpa mempengaruhi penggunaan sistem penyimpanan lama; format pengkodean baru tanpa kehilangan konten informasi. Prototipe NARA telah menunjukkan bahwa jaringan data menyediakan kemampuan yang diperlukan, termasuk dukungan untuk replikasi; federasi berbagai sistem pengawetan independen untuk mengurangi risiko kehilangan data; manajemen latensi di seluruh jaringan area luas; dan operasi massal untuk skalabilitas. Persyaratan ini diterbitkan

sebagai spesifikasi NARA Electronic Records Archive (ERA) sebagai bagian dari proses Implementasi Vendor.

Berdasarkan kriteria Audit dan Sertifikasi Repositori yang Dapat Dipercaya, kami sekarang berada pada titik di mana kami dapat menentukan kriteria penilaian, kebijakan manajemen, dan operasi standar (layanan mikro) untuk lingkungan pelestarian yang dapat diverifikasi dan mengukurnya menggunakan teknologi yang ada. hari ini. Dengan proyek NARA dan SHAMAN, kami sekarang dapat menggunakan pendekatan untuk mendemonstrasikan penutupan, kelengkapan, dan konsistensi dengan cara yang dapat digunakan untuk memigrasi semua proses pelestarian (bukan hanya metadata) ke teknologi baru, sebagai berikut:

- *Penutupan,* bahwa untuk setiap layanan mikro yang dijalankan, informasi status yang diperlukan dibuat yang dapat ditanyai untuk memvalidasi kriteria penilaian;

- *Penutupan,* untuk setiap kriteria penilaian ada aturan terkait yang mengontrol validasi;
- *Penutupan,* untuk setiap aturan ada layanan mikro yang dibutuhkan;
- *Kelengkapan,* seperangkat fungsionalitas yang disediakan oleh sistem menyediakan kemampuan yang diperlukan;
- *Konsistensi,* untuk setiap peningkatan pada sistem, kami dapat memverifikasi bahwa kebijakan manajemen tetap konsisten, dan bahwa kriteria penilaian masih dapat dipenuhi.

Implementasi Vendor

Jaringan data tunggal masih memiliki elemen risiko. Lingkungan pelestarian perlu mempertimbangkan federasi jaringan data independen atau lingkungan pelestarian independen. Kami sedang mengusulkan, melalui NARA, agar kami berkolaborasi dalam demonstrasi kemandirian infrastruktur secepat mungkin. Ini akan mencakup migrasi catatan dari SHAMAN ke arsip persisten prototipe penelitian NARA, dan validasi pernyataan tentang keaslian dan integritas. Kami berharap dapat memperluas demonstrasi ini ke vendor (Lockheed-Martin) untuk pengiriman layanan NARA ERA. Bukti alternatif dapat ditunjukkan dengan menggunakan teknologi yang tersedia dari perusahaan alternatif. Dalam kasus IBM, misalnya, ini akan terdiri dari arsip HPSS atau lingkungan penyimpanan pita serupa; Database DB3 untuk mengelola metadata pelestarian (keaslian, integritas, informasi status infrastruktur), protokol OASIS untuk mengelola layanan; dan klien Java untuk akses. Demonstrasi yang akan membuktikan kemandirian infrastruktur perlu menyediakan infrastruktur untuk mengikat komponen bersama-sama, membangun teknologi jaringan data untuk mengelola replika, sistem pelestarian federasi, menyediakan operasi massal untuk pengelolaan data, memberikan dukungan untuk ruang nama pelestarian (pengguna, file , metadata, batasan akses).

Komponen penelitian untuk prototipe NARA disimpan terpisah dari penerapan produksi melalui vendor komersial (Lockheed-Martin), dengan hasil bahwa penerapannya terutama berasal dari kemajuan penelitian. Nilai dari program penelitian ini adalah dampak yang dapat dibuatnya dalam hal menginformasikan NARA Electronic Record Archives (ERA) tentang konsep, teknologi, dan kemampuan baru yang telah ditunjukkan dalam prototipe arsip persisten NARA. Kemampuan untuk mendemonstrasikan teknologi baru dalam skala besar sangat penting untuk memvalidasi nilai kegiatan penelitian dan penerapannya.

NARA telah mengadopsi pendekatan inkremental untuk implementasi vendor. Bagian pertama, periode kontrak dasar, diberikan kepada Lockheed-Martin pada tahun 2005 dalam kompetisi berdasarkan desain sistem dan prototipe kerja mereka. Berikut
ini, proyek akan mengadopsi pendekatan inkremental dengan kapabilitas operasi awal (IOC) yang disampaikan pada tahun 2007 dan kapabilitas operasi akhir disampaikan pada tahun 2011. Saat ini vendor berkomitmen untuk versi produksi dari teknologi prototipe NARA yang akan direvisi pada berkelanjutan selama berlangsungnya proyek. Inisiatif penelitian lanjutan adalah:

- Mengotomatiskan proses pengarsipan dan mengelola pemetaan yang konsisten dari asal, administratif, deskriptif, dan keaslian yang dikaitkan dengan ruang nama logis (University of Maryland, SDSC);
- Gunakan kisi data untuk mengelola hubungan pemetaan ini untuk mengatur hubungan yang tidak bergantung lokasi (SDSC);
- Gunakan teknologi pengawetan yang dapat menjamin penemuan, akses, presentasi, dan manipulasi dokumen atau data apa pun secara independen dari infrastruktur apa pun (Universitas Liverpool).

Prototipe NARA telah mendorong kemitraan positif antara industri dan akademisi. Beberapa teknologi penyimpanan terdistribusi yang dijelaskan dalam makalah ini sekarang sedang diubah menjadi versi perusahaan untuk digunakan oleh perusahaan seperti EMC, Rolls Royce, dll. Yang berfungsi sebagai contoh transfer teknologi ke penyedia komersial.

Relevansi Proyek

Kebijakan manajemen NARA relevan dengan sejumlah besar inisiatif terkait yang meminta penggunaan teknologi jaringan data untuk mendukung federasi lingkungan pelestarian. Ini termasuk:

- **Proyek NDIIPP Perpustakaan Kongres; NSF Chronopolis.** Proyek-proyek ini meminta penggunaan teknologi jaringan data untuk mendukung federasi lingkungan pelestarian [44].

- **Perpustakaan Digital Sains Nasional NSF.** SDSC mendukung arsip tetap dari materi pendidikan yang diambil melalui penjelajahan web dari konten yang terdaftar ke dalam repositori NSDL. Arsip persisten NSDL adalah aplikasi NARA - teknologi pelestarian yang didanai untuk mendukung koleksi pendidikan tertentu. Universitas Liverpool Cheshire dan teknologi Multivalent menyediakan sarana untuk menganalisis data [45].

- **NSF Teragrid.** SDSC menerapkan NARA - teknologi pengawetan yang didanai untuk mendukung lingkungan replikasi untuk kumpulan data ilmiah lokal. Aplikasi tersebut membutuhkan pengembangan antarmuka SRB tambahan untuk berinteraksi dengan sistem penyimpanan arsip unik yang digunakan oleh Teragrid, SDSC, dan National Center for Atmospheric Research (NCAR) [46].

- **Program NSF DIGARCH (Digital Preservation and Lifecycle Management).** Baik SDSC dan University of Maryland berpartisipasi dalam demonstrasi pelestarian jenis koleksi multi-media tambahan. Di setiap proyek, koleksi tertentu disimpan [47].

[44] http://www.digitalpreservation.gov/
[45] http://www.dli2.nsf.gov/
[46] http://www.teragrid.org/
[47] http://www.sdsc.edu/srb/projects/digarch

Secara lebih umum, permintaan untuk manajemen data yang berorientasi pada aturan terjadi di banyak disiplin ilmu, serta dalam berbagai komunitas manajemen data, termasuk:

- Grup Hierarchical Data Format (HDF) untuk mendukung manipulasi HDFv5 - kumpulan data ilmiah yang dikodekan [48];
- Komunitas Open Source Network for a Data Access Protocol (OpenDAP), yang menggunakan standar terpisah untuk mengurai sintaks dan semantik data ilmiah [49];
- Komunitas infrastruktur middleware DataCutter yang menerapkan filter aliran untuk memanipulasi data ilmiah [50];
- Komunitas Data Format Description Language (DFDL) yang representasinya dapat digunakan untuk mengkarakterisasi struktur data dari komunitas di atas [51].

Karena komunitas eScience tertarik pada format data yang terdefinisi dengan baik dan semantik yang terdefinisi dengan baik, kami harus dapat memetakan aplikasi - alat analisis khusus untuk teknologi NARA dan SHAMAN. Di masa depan, pendekatan kami adalah memilih komunitas dan mendukung panggilan perpustakaan mereka untuk mengakses data. Sekarang ada beberapa komunitas eScience yang terorganisir dengan cukup baik untuk melakukan ini:

- *Astronomi IVOA:* layanan web untuk memanipulasi file Sistem Transportasi Gambar Fleksibel (FITS) yang berinteraksi dengan katalog [52];
- *Ilmu Kognitif:* Berbagi data subjek manusia yang dikontrol akses;
- *Ilmu samudra:* Mekanisme OpenDAP untuk secara eksplisit memanipulasi format data terdaftar untuk mengekstrak data fisik. Ini memberikan dukungan untuk data sensor waktu nyata, termasuk didistribusikan di berbagai fasilitas penelitian [53];
- *Sistem Pengamatan Bumi NASA (EOS):* Perpustakaan HDF5 menyerukan manipulasi struktur data untuk NASA [54];
- *Kisi data eScience Inggris Raya:* spesifikasi koleksi - tergantung disposisi dan pilihan versi [55];
- *Perpustakaan digital DSpace:* Manajemen hak dan kriteria penilaian penyimpanan digital tepercaya [56].

Teknologi NARA dan SHAMAN memiliki relevansi yang semakin meningkat untuk jangkauan media yang lebih luas, misalnya domain televisi dan telekomunikasi yang memerlukan siklus hidup pelestarian untuk menyatu secara mulus dengan produksi konten melalui penggunaan alur kerja yang mengotomatiskan aksesi, deskripsi, organisasi, dan pengawetan berbagai jenis media.

[48] http://hdf.ncsa.uiuc.edu/
[49] http://www.opendap.org/
[50] http://datacutter.osu.edu/
[51] http://forge.gridforum.org/projects/dfdl-wg/
[52] http://www.ivoa.net/
[53] http://www.opendap.org/
[54] http://eospso.gsfc.nasa.gov/
[55] http://www.e-science.clrc.ac.uk/
[56] http://www.dspace.org/

Ringkasan

Teknologi yang dibutuhkan untuk implementasi prototipe NARA dan SHAMAN mengasumsikan terciptanya lingkungan pelestarian generik. Infrastruktur ini dihasilkan melalui integrasi kemajuan dalam perpustakaan digital, jaringan data, dan komunitas arsip yang persisten. Pengembangan dan implementasi teknologi dalam pengujian arsip persisten fokus pada pengelolaan evolusi teknologi, daripada hanya mencirikan asal catatan melalui informasi representasi. Pendekatan ini terutama didasarkan pada penggunaan kisi data iRODS dan model objek digital Multivalent dengan dukungan tambahan untuk layanan perpustakaan digital untuk mendukung penemuan dan analisis data. Kemandirian infrastruktur yang sebenarnya menyiratkan bahwa salah satu komponen di atas dapat ditukar dengan padanannya tanpa kehilangan keasliannya. Panjang - tujuan jangka adalah teori manajemen data "engingeering". Kami menguraikan kapabilitas yang diperlukan menjadi sekumpulan layanan mikro, menguraikan kebijakan manajemen menjadi sekumpulan aturan, menguraikan kriteria penilaian menjadi kueri pada informasi status persisten. Dengan adanya ketiga ruang ini, kami kemudian menunjukkan bahwa pemetaan antara ketiga ruang tersebut dalam hal operasi aktual adalah lengkap, konsisten, dan tertutup. Teknologi yang digunakan untuk memberikan kemandirian infrastruktur harus dapat diterapkan pada jenis sistem dan domain manajemen data lainnya. Pekerjaan yang telah kami lakukan hingga saat ini akan memungkinkan kami di masa depan untuk fokus pada komunitas eScience dan cyberinfrastructure yang memiliki - format data yang ditentukan dan dengan baik - semantik yang ditentukan.

Ucapan Terima Kasih

Penelitian ini disponsori oleh Joint Information Systems Committee (JISC), National Science Foundation (NSF), dan National Archives and Records Administration (NARA) selama periode 1999-2007. Pengakuan penuh syukur diberikan kepada Reagan Moore dan rekan-rekannya di unit Komputasi Intensif Data SDSC di Universitas California, San Diego; Robert Chadduck (NARA); Thomas A. Phelps (pengembang Multivalent); Ray R. Larson (UC Berkeley); kolega di Universitas Liverpool (Robert Sanderson, Clare Llewellyn, John Harrison, Fabio Corubolo, Catherine Smith).

Referensi

Arcot, R., Moore, R., Berman, F., & Schottlaender, B. (2005). “Pelestarian Digital Manajemen Siklus Hidup untuk Koleksi Multi-media ”, di *Perpustakaan Digital: Menerapkan Strategi dan Berbagi Pengalaman,* hlm. 380-84. Catatan Kuliah Ilmu Komputer, Vol. 3815/2005. DOI 10.1007 / 11599517.

Altintas, I., Barney, O., & Jaeger-Frank, E. (2006) Dukungan pengumpulan asalnya di Sistem Alur Kerja Ilmiah Kepler. Di *Provenance and Annotation of Data, International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, 3-5 Mei 2006*. Diakses 4 Desember 2007 dari http://www.ipaw.info/ipaw06/proceedings/CameraReady_s5_2.pdf

Boisvert, RF, Tang, PTP (Eds.). (2001). Arsitektur perangkat lunak ilmiah, IFIP TC2 / WG2.5. *Konferensi Kerja tentang Arsitektur Perangkat Lunak Ilmiah, 2-4 Oktober 2000, Ottawa, Kanada, Prosiding Konferensi IFIP, vol. 188,* Kluwer.

Gladney, HM (2002). Perspektif tentang informasi yang dapat dipercaya. *Dokumen Digital Kuartalan, vol. 1, No. 3.* catatan 9. Diakses 4 Desember 2007 dari http://home.pacbell.net/hgladney/ddq_1_3.htm#_edn9

ibid. note 11. Diakses 4 December 2007 dari http://home.pacbell.net/hgladney/ddq_1_3.htm#_edn11

Hedstrom, M. (1991) Memahami incunabula elektronik: Sebuah kerangka kerja untuk penelitian tentang catatan elektronik (UEI), dikutip dalam Peta Jalan Riset Digital Preservation Europe (DPE), Proyek no. 034762, Bagian B1. *Pengarsip Amerika 54, 3 (Musim Panas 1991):* 334-54 (Cox).

Jeffrey, S., & Hunter, J. (2006). Mesin pencari semantik untuk Storage Resource Makelar. *Workshop Semantic Grid ke-3GGF16, Athena.* Diakses 4 Desember 2007 dari

http://www.itee.uq.edu.au/~eresearch/projects/dart/workpackages/si3.php
http://www.itee.uq.edu.au/~eresearch/projects/dart/outcome/semanticsrb .php

Larson, R., & Sanderson, R. (2006). Cheshire3: Mengambil dari berbasis grid berskala tera perpustakaan digital. *SIGIR 2006:* 730). Diakses 4 Desember 2007 dari http://portal.acm.org/ft_gateway.cfm?id=1148343&type=pdf

Larson, R., & Sanderson, R. (2005). Perpustakaan digital berbasis grid. *JCDL 2005,* 112-113. Diakses 4 Desember 2007 dari http://ieexplore.ieee.org/xpls/abs_all.jsp?arnumber=4118524

Ludascher, B., Marciano, R., & Moore, R. (2001a). Pelestarian data digital dengan self-validating, self-instantiating arsip berbasis pengetahuan. *ACM Sigmod Record, Vol 30, Edisi 3,* hlm. 54-63. Diakses 4 Desember 2007 dari http://portal.acm.org/citation.cfm?id=603876

Ludascher, B., Marciano; R., & Moore, R. (2001b). Menuju validasi diri arsip berbasis pengetahuan. Masalah Penelitian di Data Engineering, 2001, hlm.9-16. Dalam *Masalah Penelitian di Data Engineering, 2001. Prosiding. Lokakarya Internasional Kesebelas tentang. ,* hlm. 9-16. ISBN 0-7695-0957-6. DOI: 10.1109 / RIDE.2001.916486. Diakses 4 Desember 2007 dari http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=916486

Moore, R. (2003). Persyaratan konsistensi umum untuk data grid, perpustakaan digital, dan arsip tetap. Dikirim ke 12th Konferensi Komputasi Terdistribusi Kinerja Tinggi, Seattle, WA (Juni 2003). Diakses 4 Desember 2007 dari http://grid.lbl.gov/GPA/GGF7_Data_Consistency.Word95.pdf

Moore, R. (1999). Arsip persisten untuk pengumpulan data. Kemitraan Nasional untuk Advanced Computational Infrastructure (SDSC). Dikirim pada tahun 1999 sebagai bagian dari Lokakarya Arsip tentang Standar Penyerapan, Identifikasi, dan Sertifikasi (AWIICS). Diakses 4 Desember 2007 dari http://nost.gsfc.nasa.gov/isoas/awiics/

Moore, R. (2006). *Penelitian tentang arsip persisten.* Laporan Teknis SDSC 2003-06. Diakses 4 Desember 2007 dari http://www.sdsc.edu/NARA/Publications/data-preservation.doc

Moore, R, dkk. (2006). *Sistem pengetahuan berbasis kendala untuk grid, digital perpustakaan, dan arsip persisten.* Laporan Teknis SDSC 2005-9.

Moore, R. & Marciano, R. (2007). Lingkungan pelestarian prototipe. Kertas dikirim ke *Tren Perpustakaan.*

Moore, R., Arcot, R., & Marciano, R. (2007). Menerapkan Digital Tepercaya Repositori. Diakses 4 Desember 2007 dari http://ww.ils.unc.edu/digccurr2007/web-abstracts/moore_abstract_6-4.pdf

Moore, R. & Smith, M. (2007). Validasi otomatis dari repositori digital tepercaya kriteria penilaian. *Jurnal Informasi Digital,* Vol. 8, No 2 (2007). Diakses 4 Desember 2007 dari http://dspace.mit.edu/html/1721.1/39091/Moore-Smith.htm

Moore, R., Baru, C., Arcot, R., Ludaescher, B., Marciano, R., Wan, M., dkk. (2000, Maret). Arsip digital persisten berbasis koleksi. *Majalah D-Lib, 6 (11).* Diakses 4 Desember 2007 dari http://www.dlib.org./dlib/march00/moore/03moore-pt1.html

Phelps, TA, & Watry, P. (2005). Arsitektur tanpa kompromi untuk dokumen digital kelestarian. *ECDL,* 2005: 266-277. Diakses 4 Desember 2007 dari http://multivalent.sourceforge.net/Research/Live.pdf

San Diego Supercomputer Center (SDSC). (1999) Laporan Teknis DOCT, Bagian 2.4. Diakses 4 Desember 2007 dari http://npaci.edu/DICE/Pubs/replication.doc

Smorul, M., JaJa, J., Wang, Y., & McCall, F. (2004). *PAWN: Produser-Arsip Jaringan Alur Kerja untuk mendukung pelestarian digital.* CS-TR-4607, UMIACS- TR-2004-49. Diakses 4 Desember 2007 dari http://umiacs.umd.edu/research/adapt/papers/UMIACS-TR-2004-49.pdf

Smorul, dkk. (2002). Metodologi antarmuka produser-arsip: Standar abstrak, Komite Konsultatif untuk Sistem Data Antariksa. CCSDS-651.0-R-1, Buku Merah, Desember 2002. Diakses 4 Desember 2007 dari http://public.ccsds.org/publications/archive/651x0b1.pdf

Thibodeau, K. (2001, Februari). Membangun arsip masa depan. *Majalah D-Lib, 7, (2).* Diakses 4 Desember 2007 dari http://www.dlib.org./dlib/february01/thibodeau/02thibodeau.html

Watry, P., & Larson, R. (2005). Buku putih Cheshire framework: Menerapkan support for digital repositories in a Data Grid environment. In *Proceedings of the IEEE Conference on Globally Distributed Data (2005),* pp. 60-64. Retrieved December 4, 2007 from http://cheshire.berkeley.edu/Cheshire_Sardinia.pdf