Meeting Notes

Advanced Camp June 19, 2008

Program and Presentations

Introductions and Goals of the Workshop

*** R.L. Morgan, Senior Technology Architect, University of Washington**

After attendees introduce themselves and share their interests, we'll review the issues and questions that bring us together, with a view toward outcomes: potential initiatives and focused research into requirements and solutions for collaborative academic services.

Boundaries are becoming very fuzzy. Collaboration is increasingly happening between organizations, and larger numbers of organizations are participating in the various projects.

Serendipity is important, and we don't want security to be a barrier to innovation and getting our work done.

Metadata issues - naming assets seems to be a key element of the problem. Linking metadata to assets as they move around, and persistent naming, are important as this grows.

Q: to what degree has your institution embraced the notion that scholarly content is the responsibility of central IT?

Increasingly, the scholarly realm is more independent and thus are leading central IT into new areas.

Q: Why would academics want to store their content in central repositories? What is the value proposition for them?

Enhancing exploration and productivity within disciplines is a central goal...

Q: How do digital library assets factor into the broader equation?

CI Salad: Defining the Problem

What needs must cyberinfrastructure for the humanities meet? Two Mellon-funded projects, SEASR and Bamboo, are trying to determine the leading problems to be solved to support digital humanities efforts and the frameworks, tools, and services needed to solve them. This session will brief us on what has been learned up to this early point in these projects, the approach each is taking, and the relationships they'd like to develop between their efforts and others.

*** Loretta Auvil, Senior Project Coordinator, University of Illinois at Urbana-Champaign**

SEASR
How can we leverage technology infrastructure framework to support multiple domain apps?

Q: Who would be building custom apps in this environment?
A: Combination of domain experts and developers. Java, Python, and GLisp are used. End users are the target audience for these apps, thus UI is very important. The ability to explore is also key, to enable users to see how others are using the data.

Q: Are mashups supported in the SEASR/Meandre framework?
A: Transformers can be written to translate data if required, or XML is the common lingua franca. Components from multiple repositories can be combined as appropriate. Licensing issues sometimes can be sticky. Exposing what is happening, and provenance capabilities, are key to making this useful.

Q: Will SEASR supply actual services that can be used to build things?
A: This is flexible, current focus is on building repositories where developers can register their component libraries. Eventually the goal is to provide services that are more baked for users.

Various navigation, visualization, and exploration schemes are being supported, e.g. tag clouds.

Q: Is this too overwhelming for arts and humanities folks?
A: Goal is to make this as user-friendly as possible, but part of the challenge is increasing general understanding among those communities about what is possible and thus spurring their thinking about ways in which this could be useful to them.

*** Steve Masover, Data Architect, University of California, Berkeley**

Project Bamboo

Fuzziness/serendipity are key in Arts and Humanities -- key to useful discovery.

Provenance, and how and by whom resources are used/cited, is also very important. "Chains of credit" can be very important in tenure evaluations, as an example... IPR concerns in the commercial sector are an interesting parallel to this. There will likely also be a privacy-related tension in this - scholars want to know who is using their data, but don't want their use to be monitored.

Q: Data privacy, and regulatory concerns - how are these addressed in exposing datasets? How does the infrastructure support this?
A: This is on their radar, but too early for answers just yet...

Q: Teaming in Arts & Humanities seems to be rare, but digitizing content seems to be a big and growing issue. Is there a culture gap here?
A: Workshop participants are self-selecting, and they tend to be more team-oriented than perhaps their colleagues may be.

Service-Oriented Projects in Higher Ed

Several community source service-oriented efforts address issues particular to higher education. In this session, we will explore these efforts and hear about the experiences of developers and early adopters.

*** Daniel Davis, Chief Software Architect, Fedora Commons**

Fedora Commons is driving towards being a more community-driven software development effort. Integrating capabilities from many other sources. Meeting with DSpace to look at collaboration opportunities... Looking at long-term sustainability for the project, and organizations doing similar and complementary work.

Trending towards economies of scale - multiple institutions, integrating with commercial products to support more complex implementations.

Whose responsibility is the curation of materials ongoing? Information lifecycle issues, especially for new information that is created in digital form. These are increasingly looking less like "documents" i.e. you need external software to view the material... How will this be funded/supported ongoing? What is the value proposition? Decoupling the services is good for preservation. Format migration becomes adaptation to new software.

One person's metadata is another person's data - it all needs to be linked together and made accessible.

Q: Is it possible to federate Fedora Repositories? What capabilities are there for supporting distributed management?
A: Yes, you can use anything as an identifier, thus you can select the naming scheme that makes sense for you. Plus you can have alternate names for objects. You can deploy an enterprise search engine that will search on your criteria, and return objects wherever they may live.

IPR- Jim Henson site has a repository-centric security model. Muradora project moves policy driven operation (based on XACML) up into middleware. Any system that depends on an app on a system you don't control is by its very nature a problem for the protection of IPR. Need a well-defined trust and security model.

"Vendor-driven architectures" tend to be problematic... Watch out for vendor lock-in, is the term "standards-based" subject to interpretation by vendors?

Q: What is the overlap between users and institutions that want to federate content, and the identity federations we are building?
A: They are complementary capabilities, need to interoperate with each other. Duplication of information is permitted, e.g. person object associated with an account, also a pointer to an element of scholarship. How might these be merged?

Q: What is anticipated timing if institution wants to setup a federated repository and ensure it is reusable and extendible?
A: This is being done now if your initial expectations are modest, more work happening on this including in the commercial space.

They don't want to write more code than necessary, nor recreate all of the testing etc. that others are doing.

*** Jens Haeusser, Director, Strategy, The University of British Columbia**

Kuali Student

Q: To what degree are functional requirements being driven by academics v. admin?
A: We are seeking a balance, trying to reflect the broad range of the community, reflecting a broad range of practices.

Learning unit management is an early priority, treating them similar to SKUs for flexibility.

One mark of a good ESB (Enterprise Service Bus) in this context is the ability to integrate easily...

Q: Are there use cases where services are outside the enterprise? How are these being addressed?
A: Sometime from an outsourced vendor, sometimes from another campus. Sometimes there is interesting information supporting your business processes, but which are outside of your organization.

Q: The Educational Community License (ECL) is close but not identical to Apache license, is that a problem? How is this reconciled?
A: Contributions to other technologies are done using their licenses.

Computing as a Service

What tools and frameworks do we develop to make research computing into a service-oriented endeavor? The OGSA and several cloud computing vendors are developing standards to create a grid-like architecture based on web services. This session will explore the current state of cloud computing and how higher ed can use these services.

*** David Gimpl, Software Engineer, IBM Corporation**

IBM Blue Cloud Initiative

Question to ask yourself when considering cloud computing - What is the tolerable outage time that is acceptable for your key application(s)?

Questions to ask the vendor - How fast can a VM image be created initially, and then provisioned into production when more capacity is required?

Q: Smearing boundaries between HPC cluster and retail compute clusters...

Ken Klingenstein: note Dennis Gannon's paper - a computational data center - a science cloud - 3 different scientific approaches where shifting from a grid to a cloud is the right choice.
http://www.extreme.indiana.edu/~gannon/grid-article.pdf

*** Mark Morgan, Research Faculty, University of Virginia**

Genesis II
http://www.cs.virginia.edu/~vcgr/
http://vcgr.cs.virginia.edu/genesisII

Q: Open Grid Forum (OGF) is potentially in a period of transition, what happens to the standards if OGF goes away?
A: Some may fade into oblivion, but others are being picked by vendors and thus will live on (e.g. BES used by IBM)

Q: OASIS is starting up a web services harmonization activity. Would that be a logical home for this work?
A: Traditionally organizations like OASIS have focused more on web services than on grid-related services, which are more about "the whole package." OASIS would likely not be interested in taking these on.

Q: What about a user who wants to put up their own service?
A: Job Submission Description Language (JSDL) can be used to describe any job, but that doesn't necessarily mean you would give that job to a BES (Basic Execution Service) container.

Q: How does a client know what services are available, and how to get to them?
A: We are still developing a solution for this. Users are typically associated with a particular grid, and thus can be informed about what is available, but for other grids this is not so straightforward.

Q: Is there an impact on performance from all of this layering? What is the overhead, is it all up front?
A: Morgan) There is overhead ongoing, a lot of it based on I/O. they are trying to get past this problem, but they are passing XML over HTTP which by its nature is not very efficient. Note the distinction between HPC and high throughput.
A: Gimpl) Metrics and instrumentation are key. Why are there performance issues associated with a particular segment? BlueGene is both HPC and high throughput, but serves different needs.

Q: Note utility characteristics - how does this factor in?
A: You would be seeking cost/benefit answers for a particular job, depending upon your priorities.

Note that there are other specs coming out of OGF for the aggregation of metadata about services. Generally you know the characteristics of the service required to meet your needs for a particular task.

Q: Deployment models - organizations tend to want to outsource things they understand well, since they are then able to evaluate the performance. Are we moving to a model in which our users would not know or care where an app lives. Are the migration capabilities in place to support this?

A: Gimpl) IBM supports a utility model, billing you for the services you actually use. The organization decides what services to use them for. Some organizations start apps and services internally, then migrate them to the cloud when it becomes in their interest to do so, e.g. when it is more efficient and cost effective to run them in the cloud. There are economies of scale that come into play in the utility computing model that are very compelling.
A: Morgan) the university has to maintain a certain amount of infrastructure, some of which is not continually in use, and some researchers don't have the budget for clusters but still have needs. Thus it makes sense for the university to utilize this excess capacity to support these users when it makes sense to do so.

Q: Is UVa using its end-user PKI for this?
A: Yes, but not as the sole authentication factor.

Panel: ESBs and Widely Distributed Services

*** Session moderator: R.L. Morgan, Senior Technology Architect, University of Washington**
Many large organizations use enterprise service bus (ESB) products to integrate services. While loosely coupled technically, these tend to be single organization and centrally managed deployments. A panel of ESB experts will reflect on the applicability of technology and methods to multi-organization, decentralized, serendipitous, project-centric service environments.

*** Nigel Watling, Technical Evangelist, Microsoft Corporation**
Cloud Computing and the Internet Service Bus
Proprietary, highly secure apps are obviously not good candidates for cloud computing. These belong on your premises as a general rule.

"Microsoft is very much committed to open standards."

It is not unforeseeable that an organization could have services running in multiple clouds, and it is essential that they be interoperable.

Q: In terms of standards for pub/sub and service type messaging, we are not really there yet. Not many are widely viewed as industry standards. What protocols are you supporting?
A: Any that we use will have to be open source and open standards. We have heard our customers insist that their vendors must be interoperable.

It makes sense to use cloud computing to take the support load off of internal IT staff, if the economics make sense and the SLA is adequate for our needs. The goal is to make the fabric as simple as possible to connect to. Utility is the model we are going after.

Q: What do you mean by "firewall-friendly messaging"? what firewall are you referring to?
A: Callbacks to a client can be complex, and enabling it to traverse firewalls can be a challenge. WCF (windows communication foundation) infrastructure is an approach we have developed for this. IT admins still have the ability to permit deny access, but we have developed solutions for the hard coding problems

Q: What about PII - HIPAA, FERPA, etc. - issues around data moving outside the organization?
A: This needs to be worked out over time, not a simple solution... Some apps and data just don't make sense to put in the cloud, and it is up to the organization to make that decision.

*** Roland Hedberg, Internet Architect, Umea Universitet**
Open Metadir (OM2)

Swedish universities are using a common software infrastructure, and are currently sharing information between them about students moving around to take courses.

DNS SRV records are akin to MX records, standard method of locating the current instantiation of a particular service.

*** Brian Busby, Collaborative Apps Manager, University of Wisconsin-Madison**

They are just beginning their ESB implementation project, and thus they have more questions and answers right now.

Q: Your ESB selection was largely based on financial considerations. Are there capabilities not available in Cape Clear?
A: Transform, orchestration, and transport were the 3 key capabilities we were looking for, and Cape Clear has all of these.

Q: What was the process that the state of Wisconsin went through in selecting Cape Clear?
A: mostly state department talking among themselves...

Q: Have you done simulated loads or tests to prove that performance is adequate? What tools do you use?
Empirics is the tool we are using

Q: How do ESBs relate to metadirectories?
A: Cape Clear does not have a registry for defining services. Metadata is in the payloads, it really is just a broker for service calls.

Q: What changes when you switch to ESBs from a silo app?
A: The application can no longer control all that it used to, thus it really just needs to be able to talk services and rely on the ESB to do what it is supposed to do. A lot of it has to do with app owners getting comfortable with the concept and learning that it is reliable. It is important to start small, select apps that are aligned with your business processes and represent pain points.

Q: When querying PeopleSoft, how are you getting data out?
A: There are perceived performance issues related to direct access to PeopleSoft Student. They do a single extract of the tables they need, into an operational data store, and work from that.

Discussion Groups: Data Models, Governance, Service Discovery

As higher education moves from a centralized, single-institution model of enterprise information to a more service-oriented, federated approach, we will need to alter our data models, reconsider our governance, and improve service discovery. In this session, we will break out into three groups (data models, governance, and service discovery) to discuss how each of these is disrupted by these new approaches and what we can do and have done to mitigate disruptions.

Data Models
Facilitator(s)

- *** Roland Hedberg, Internet Architect, Umea Universitet**

Is formal data modeling important enough that taxonomies and controlled vocabularies should be put in place before beginning to code? What development methods lend themselves to adaptable data models? Who needs to play together in this space, both inside and beyond campuses?

Q: How do you deal with different names for a particular object or service? Different naming conventions...

What about users with multiple affiliations? How can/should affiliations be represented? A URI expressing the organization and department?

IPED codes were suggested as a potential option...
"facstaff" and "student" are common entries, and not very helpful

What is a "student"? the definition differs when you move between applications. E.g. student housing uses it to refer to someone allowed to rent a space.

Is the representation and discoverability of services also covered under this umbrella? Yes...

Getting 802.1x working across organizations is a problem, but so is getting it to work right within organizations.

In significant ways this boils down to knowledge representation...

What standards cover the representations of values?

Different SPs want affiliation represented in different ways. How does an IdP accommodate this complexity? How are multiple affiliations handled?

Governance
Facilitator(s)

- *** James Leous, Manager/Research Programmer, The Pennsylvania State University**

Who owns data in a service-oriented model? Can the role of data steward evolve to deal with federated data, or is something new needed? Are federations (for example, InCommon) sufficient to work out data-sharing agreements, and if not, who else should be involved? What types of people are "around that table" for such discussions at your institution?

Service Discovery
Facilitator(s)

- *** Thomas J. Barton, Senior Director for Integration, University of Chicago**

What means are used to facilitate service discovery on campuses? Can they extend or articulate with federated services? How are service gaps or overlaps identified and dealt with? Who needs to play together in this space, both inside and beyond campuses?

Friday 20-Jun-08

Discussion: Privacy and Policy

This session will discuss the privacy and policy issues loosely coupled and distributed environments create or illuminate, such as sharing operational policy and procedures across sites, building trust among organizations to share and manage well-run services, and maintaining appropriate privacy policy.

*** Merri Beth Lavagnino, Chief Information Policy Officer, Indiana University System**

How are distributed services projects affected by privacy and policy?
View them as enablers rather than barriers...

Not all policies are formalized and documented, but this helps when there are disagreements or differences in interpretation.

"Reasonability" may be a challenge when you cross institutional boundaries.

Set ground rules - delineate areas of responsibility and guide decisions.

Sometimes it may be hard to interpret what parts of which policies apply to particular actions or tasks. In those cases it is best to consult your interpreting body or person, wherever that role resides within your organization.

Q: How does Google analytics figure into data collection issues?
A: Indiana U. does not permit this since the data lives outside their control. When central IT doesn't provide an adequate solution, users go elsewhere, and this is a good example. We are pushing central IT to provide this service.

Q: Federation level policy v. that of individual institutions... at what point does a federation look like a 3rd party? What is the trigger?
A: Whenever you don't have direct control over the data, separate legal entity.
A: This is not the federation operator, which is not involved in the transactions.

Removing barriers to collaboration is our common goal, and removing and clarifying policy barriers to the extent practical is an important part of this. Being able to operate under clear, well defined IPR and privacy infrastructures is essential...

*** Session moderator: Kenneth J. Klingenstein, Director, Internet2 Middleware and Security, University of Colorado at Boulder**

International considerations are a significant factor.

The lack of a strong set of federal laws had caused the states to develop laws on their own, which makes cross-state interactions challenging at times.

EU priorities generally are to not go for consent unless absolutely necessary because it is very hard to obtain.

EU countries are not treating non-EU countries as if they were in the EU for the purposes of their privacy policies.

Notice, choice, and access security are common threads...

EU: When you cease your engagement with a student, you must erase all logs associated with that student. But a student's affiliation normally doesn't end when they graduate, i.e. alumni relationships...

Q: Do these policies apply to EU citizens in the US taking classes?
A: Unknown...

EU is currently working to define and triage (and normalize) what is to be considered PII under their regulations.

Are IP addresses considered PII in the EU? Because some are static, and can be linked with an individual, unless you know you are dealing with a dynamically assigned address you must treat it as PII.

Is eduPersonTargetedID (EPTID) considered PII in the EU? It is persistent, non-reassignable, can be different for every site you visit so as to avoid correlation between sites. Thus we are hopeful that it will not be considered PII.

In the absence of specific laws covering your situation, if you review and apply the core principles to guide your decision making, that would be considered adequate due diligence.

OASIS TC: Cross-enterprise Security and Privacy Authorizations (XSPA)
http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xspa

Is "consent" considered a new service? Does every service need to be "consent aware"?

Q: Are the EU privacy directives a lever we can use with industry, to persuade them to do federated access?
A: It certainly can't hurt... the EPTID issue (i.e. is it considered PII?) will likely be a key driver. Many other countries follow the EU model, and many other countries obviously do business in the EU and thus need to get on board.

Also China is starting to flex its muscles, and to date has little concern for privacy.

The US has sector-specific privacy laws, but not overarching privacy regulations.

Note the IETF initiative on trust, identity, and the Internet which is trying to bring trust and identity back into the RFC process.
http://www.isoc.organization/isoc/mission/initiative/trust.shtml

What do you do when there it not a trusted IdP in your country? OpenID assertions are dubious, by definition. Identity will likely start to mingle with more protocols, out of necessity.

Q: What is the timetable on EU decision?
A: Likely 3Q2008, there will be broad coverage when it happens.

Discussion Group Report-Outs and Lightning Talks

* **Session moderator: R.L. "Bob" Morgan, Senior Technology Architect, University of Washington**

As a follow-up to the previous day's discussion groups, attendees will provide short reports of their breakout session. We'll then move into Lightning Talks. Do you have a practice or interesting approach to share? Or would you like to connect with someone with a similar challenge and collaborate on a solution? These very short talks will provide a final chance for attendees to discuss a good idea or opportunity for peer networking.

Data modeling: (RL "Bob")
Modeling organization data is difficult, the world is complex and increasingly hard to label and categorize...

Service Discovery: (Liz W - UMn)
What would a service description contain?

DNS SRV records work well, but are rarely used...

XRI and XRDS enable resolution services...

Governance: (Jim Leous)
Domain governance - governance largely revolves around organization domains, and data attributes and who controls it.

Who owns the data in a service model? How is the data stewardship model changing? People, apps, and services are all requesting data. How is consent obtained and expressed?

How to keep track of directory of services available, and who can use what? How are notifications handled when services change?

This is all evolving rapidly... helpful to survey domains in need of governance.

Lightning Talks

Rob Carter, Duke
Misplacing IP - Artifacts are escaping our spheres of control. How do we track where they are going?

IP Provenance - how do others know that artifacts are genuine or derived or modified?

Digital signatures are one possibility, add to metadata for artifacts? Create an authoritative registry?

Q: How does correlate with creative commons efforts to attach licenses to content? Users can search for CC licensed data to use...

When an object is transformed by a service, is it then signed again? When there are multiple versions of an objects, which one is current.

Loretta Auril, UIUC
Workbench for data analysis... classifying data and sending back to web app via an xml file.

Scotty Logan, Stanford
IAM and well-behaved apps

For reference: http://oauth.net/

Wrap-Up and Findings

* **Session moderator: Kenneth J. Klingenstein, Director, Internet2 Middleware and Security, University of Colorado at Boulder**

The program committee will lead a discussion session to summarize the final points from the last day and develop any conclusions and next steps.

Tom Barton
This is a large and somewhat ambiguous topic area...

RL "Bob" Morgan
Project Bamboo best example of a central activity that is trying to determine how to support scholarship - asking scholars what they actually want...

Often interactions with scholars have baggage from past interactions, and are constrained by the available services central IT can offer. Increasingly they are looking for us to advise them as they evaluate external service offerings, rather than directly support them with centrally provided services.

What are the policy and IPR issues around using these external services, when data leaves our direct control?

SEASR and Bamboo are academic services, but more focused on supporting researchers in all of their individuality.

SEASR goal is to provide tools users can use to solve their problems. Bamboo provides central authority, building tools to help create a learning community.

We need to build services that attend to both of these needs...

Ken Klingenstein
This has been a very rich and overwhelming meeting. It is hard to see the forest for the trees...

Did we do the cross stitch on registering, discovering, and using services? Is there holism? Are generalizations possible in this space?

Our definitions of services are often wildly different - grid services, Bamboo world of services providing point specific solutions.

Is the spectrum of services so broad that we can't say anything universal?

Yes, but... it does seem that there is little or no commonality, but hard for universities to support scholarship without finding commonality.

What is the service, how/where do I get it, are there different instances or versions of the service and if so how do I choose.

Taxonomy/classification needed, how to distinguish distinct parts and then generalize upward? How do we apply adjectives? What framework can we use to understand the space, especially as so much happens outside the scope and control of the institution. How much sub-classification is appropriate or needed?

Meta-services?

Machine oriented services (e.g. grid) are just one piece of the puzzle. Human oriented services are another facet. What others? Is overkill a danger if we go too deep in trying to classify?

Local and hosted is a key distinction. How far does "local" extend?

Loosely coupled services - interrealm vs.. campus based a useful categorization to apply? What changes? Policy and engineering...

Library approach to metadata based on cataloging and classification. We are trying to classify things while also trying to classify and describe them - adds to the complexity.

Instance vs.. the abstract. First we need to discover that there is such a service in the world, then determine which instance I want to use. Item record vs.. a bibliographic record, in library terms. Sometime I find an instance and want to go up to find others...

Web services vs. services - what analytic dimensions change?

Note OASIS "harmonization of web services activity." Web implies constraints. Different types of web services imply certain identifiers and approaches. E. g. REST vs. WS*

Mainstream technologies are mainstream because they are useful.

Firewall impacts - forcing everything over ports 80 and 443... web-based apps have become the default.

Mellon interested in opportunities to support useful R&D in this space... What would be logical areas to pursue?

Kuali library services? Community source project, service oriented, cataloging, ingestion, circulation, acquisitions, metadata, etc. This is a clear need today... searching various library databases is confusing to users - what are they actually searching?

Note that CLIR is working on open source standardized core for integrated library system.

Registration/organization - recall the vetting process Gopher used. Is there a universal registry? No, there are multiple single registration points.

We have many different communities, operating differently, no single model will apply to more than a few of them. But what common principles can we extract?

What would the meta-architecture look like? We need consensus on commonalities before diving into creating a reference architecture.

Is there useful work to be done fleshing out best practices in building and using services? Isn't that what OKI is doing?

Disciplinary communities have done a lot of work describing and classifying their areas (e.g. ACM, IEEE).

eLearning and library spaces are ahead of the research community in many ways, in terms of classifying and registering service functionality.

We need to abstract out the needs we are trying to meet (what), then let the frameworks and tools be built to meet them (how).

Many R&E processes are moving off campus whether we want them to or not, to external service providers.

How do we make our architecture flexible and robust enough to make this workable?

Concrete connections drawn?
- All who spoke about open source projects expressed desire not to reinvent the wheel.
- standard common "software stack" will enable resources to go where they are really needed.
- Common need is to store rich objects
- Identity services - everyone needs to deal with authn/z and auditing. Vertical apps rarely do this well, but we don't have a lot of good options. How do we fill this gap?
- Bamboo and Kuali are getting a lot of attention and will likely cause something to happen.

There is merit in considering a domain-centric approach to services. The complexity of a canonical service (e.g. IdM) is such that the actual service instance should be the same across multiple domains - but not a common instantiation?

Different domains require adaptations? E.g. discovery, representation, and search of services - both human and machine-readable

Chain of authenticity - provenance. Peer review process in journals is an example, but how to apply this to digital assets? Archiving, authentication, etc...

Assessment and evaluation of services - something a university ought to provide? A service unique/canonical to this realm. This will become more important as budgets decline...

We are all struggling with many of these concepts - much of this is new ground. How can we help each other?