Challenges and Opportunities for New Collaborative Science Models

Report from the AAMC Task Force on Information Technology
Infrastructure Requirements for Cross-Institutional Research
Executive Summary

The AAMC Group on Information Resources (GiR) established a Task Force on Information Technology Infrastructure Requirements for Cross-Institutional Research to provide AAMC member institutions with an overview of IT requirements in several areas deemed critical for the development of effective cross-institutional research collaborations. Increasing emphasis on multi-center studies, evidence-based medicine, translational research, and biomedical informatics requires the deployment of new technologies and policies to support modern biomedical research. The task force examined three areas relevant to successful collaboration.

1. Electronic collaboration tools to support virtual research communities.

Pursuing research requires collaboration often without regard to organizational boundaries. Standards-based tools need to be defined that will enable the effective and secure exchange of data, information, and knowledge among researchers, both within the United States as well as internationally.

2. Cross-institutional authentication and authorization.

Simplifying the processes for correctly identifying individuals (authentication) and providing them appropriate access to data and systems based on their identity (authorization) will be essential for building effective cross-institutional collaborations in the dynamic environment of modern biomedical research.

3. Policies, procedures, and standards to ensure the integrity of data and networks.

Standards and processes need to be implemented for certifying compliance and assuring trust, or the willingness of institutions to accept the increased risk of collaborative, networked initiatives will impede cross-institutional translational research.

Collaboration represents the future for basic and clinical research in academic medicine. The concept or practice of team science is not particularly novel. What is notable is that research teams are no longer confined to one institution but now span organizational boundaries, and their work is being enabled by information and communication technologies.

Information technology offers a collection of helpful tools to foster communication and data sharing. The task force examined some of the current and emerging communication and collaboration tools used to support modern biomedical research. Many of the technologies have become ubiquitously available; but few, if any, of these will likely meet institutional requirements for security, privacy, or
access control. Therefore, it is extremely important that each technology solution is assessed in the context of the particular research team involved – including the people on the team; the types of information being collected and stored; the services required; and the risks associated with the inappropriate access to data, information, results, or the workings of the group itself.

Data sharing requires more than technological interconnections between systems. Building systems in a manner that assures interoperability and data sharing is a challenge from technical, policy and personnel perspectives. Processes are the methods used to acquire, accumulate, analyze and annotate research data. Data reuse is facilitated by adopting standards that allow its meaning and context to be shared between researchers. The people who use these technologies and methods need be comfortable in the knowledge that sharing intellectual assets will not compromise their own progress.

Absence of scalable information security and organizational trust agreements limit data sharing. A well-defined common model agreed upon by the academic medical community and congruent with similar initiatives in broader academia and government will facilitate the use of national and even international shared resources. Taken together, these factors would facilitate clinical and basic research by providing an agreed-upon infrastructure for securing clinical data exchanges with the research community.

The task force makes the following recommendations:

- Leverage existing institutional collaboration tools such as learning managements systems.
- Develop common data sharing agreements within the institution that can serve as models for broader inter-organization agreements.
- Adopt and promote standards for representation of research data.
- Federate identity management methods to automate delivery of shared data and services in academic medicine.
- Create automated methods to share researcher credentials, expertise, and authorizations to expedite access to data and systems.

Together, these recommendations define a roadmap to simpler, more secure research collaborations in academic medicine.
Challenges and Opportunities for New Collaborative Science Models

Rationale for Collaborative Research Computing

The scale and complexity of today’s biomedical research problems demand that scientists move beyond the confines of their individual disciplines and explore new organizational models for team science. Research teams, particularly those that span institutions, represent the fastest growing type of authorship structure and are producing many of the highest-impact papers. [Wuchty et al., 2007; Jones et al., 2008] The concept or practice of team science is not particularly novel. What is notable is that research teams are no longer confined to one institution but now span organizational boundaries, and their work is being enabled by information and communication technologies. The promise of team science has also attracted the attention of a number of foundations and government organizations, which have been actively working to understand and promote team-based discovery. In the area of biomedical research, the National Institutes of Health (NIH) Roadmap Initiatives are encouraging investigators to find new ways of combining skills and disciplines in the physical, biological, and social sciences to realize the great promise of 21st century medical research. More recently, the Clinical and Translational Science Award (CTSA) initiative is promoting multidisciplinary and interdisciplinary clinical and translational research as a means of catalyzing the application of new knowledge and techniques to clinical practice at the front lines of patient care. The concept of the “research team” is also expanding to include a much more diverse population. Community engagement is a mandate for the CTSA initiative, and many research teams now include patients and health care professionals in addition to the traditional institutionally based biomedical and clinical researchers. The important information, insights, and connections that each of these groups brings to a research question facilitates the research process and also expedites the translation of the research findings to improved outcomes.

One of the primary reasons for the increase in team-based research has been the remarkable advances in communication and information technologies. This has truly proved to be a two-edged sword for biomedical and clinical research. New electronic tools and services that facilitate collaborative activities and foster group (social) interactions are being made available over the Internet, almost on a daily basis, often without cost. Such tools are being widely adopted by individuals for social, personal, and professional interactions. This is effectively cutting down barriers for the use of these technologies in the support of research. The other edge of the blade is that research and health care-related organizations are being held to increasingly rigorous standards for information security, privacy, and access control—issues seldom considered in the design or operation of many publicly available collaboration and communication tools.

NIH’s research strategy calls for interdisciplinary, and, at times, interinstitutional, teams of researchers to advance our understanding of biological systems, to help create the research team of the future, and to reengineer the clinical enterprise. The ability to share research and patient care information quickly and easily across organizational boundaries in a controlled, confidential, and secure manner is key to the success of this initiative. Interdisciplinary and interinstitutional, almost by definition, require solutions that facilitate the collaboration of heterogeneous units. For interinstitutional research to evolve beyond the simple exchange of data among a limited number of investigators, an appropriate technological infrastructure needs to be in place along with necessary policies and procedures that will facilitate team research and appropriately mitigate risk.
Access to collaborative tools and information resources across institutions is particularly challenging in the areas of biomedical and clinical research. This is due to information security requirements associated with the use of protected health or sensitive personal information and the desire of the individual institutions and investigators to safeguard their intellectual property and minimize risks and liabilities. To effectively control access to electronic resources, processes and technologies are needed that will reliably associate individuals with their digital identities (authentication) and individuals also need access to the resources they are permitted to use (authorization). This is currently done to an acceptable level within most organizations, providing access to various internal systems and services. The major challenge arises when dealing with cross-institutional research teams in which individuals do not have authentication credentials and other digital identities that can be recognized, evaluated, and trusted by information systems hosted by multiple organizations. This has resulted in the implementation of a variety of identity management work-a-rounds that are extremely frustrating for potential users, will not scale effectively as the number and complexity of research teams increase, and often provide little if any security.

Implementation of a common model for authentication and authorization across institutions would allow research teams working either between or within institutions to easily and effectively exchange data and research results in a secure, well-controlled fashion. The NIH as well as several large national biomedical research initiatives, such as the National Center for Research Resource’s Biomedical Informatics Research Network (BIRN), National Cancer Institute's Cancer Biomedical Informatics Grid (caBIG), CTSA, and the NIH's Genome Wide Association Study, have made great progress in establishing infrastructure frameworks and standards to facilitate sharing information among distinct research communities and are implementing Federated Identity Management as a common model for authentication and authorization of individuals. Generalizing these various approaches to promote information-sharing is not as simple as reusing the same technologies and practices. Smaller research collaborations may not have the resources to acquire or sustain the needed technical components. Furthermore, many initiatives have been built based on use cases and research data models specific to their community. Extending these to a more generalized format is often not a primary objective. To effectively support cross-institutional research teams, each organization must consider their technology infrastructure, the processes they have in place, and their staff. Appropriately structured and staffed, investigators working within different units in an institution or across different institutions need to have access to robust, flexible, and secure electronic systems that can bridge geographic distances, work across time differences, and comply with the research and technology policies of each institution.

The AAMC Research Task Force has identified, in this report, several topics and tools that academic medical centers should consider if they want to participate in the growth of interinstitutional research. As with any tool kit, the particular tools needed and the way that each tool should be used will depend on the problem being addressed at a given institution and the risks that need to be mitigated. Even with the proper technology in place, deciding how best to use the systems and what business practices to adopt is complex and will require continued vigilance in this rapidly changing environment. The Task Force presents several questions investigators and institutions must answer as they create their research collaboration environment.
Core Components of Interoperability

Blending bench research data with health care and clinical information holds the promise of rapid discovery of new treatments and cures. Clinical translational research information systems will need to connect disparate systems and combine sundry data elements to promote bench-to-bedside discoveries. Rarely, however, do research and clinical data repositories come together in ways that greatly facilitate connections among these disparate data sources. Building systems in a manner that assures interoperability and data sharing is a challenge from technical, policy, and personnel perspectives.

Processes are the methods used to acquire, accumulate, analyze, and annotate research data. Data reuse is facilitated by adopting standards that allow its meaning and context to be shared between researchers. The people who use these technologies and methods need be comfortable in the knowledge that sharing intellectual assets will not compromise their own progress. When research systems simultaneously achieve these goals, multi-institution translational research can readily advance.

Technology

Technology encompasses the individual pieces of hardware, software, databases, and data networks. To achieve interoperability between each piece, widely adopted and uniformly implemented standards knit them together so that disparate information systems can easily and unambiguously exchange data with each other. The tools described later in this report illustrate how employing common standards enable the sharing of information that fosters collaboration.

However, not all widely adopted standards have been implemented uniformly. As Hyper Text Markup Language (HTML) evolved, vendors competed with each other by enhancing their browsers in proprietary ways while adhering to HTML definitions. Web pages created with vendor-specific extensions to HTML display differently in the browser made by a competing vendor. Health Level Seven (HL-7) is a standard way to represent many kinds of clinical information. Virtually every clinical information system can send and receive patient information using HL-7. Unfortunately, organizations find it difficult to exchange all but the simplest elements of patient information between different clinical systems because of subtle differences in how HL-7 is implemented. These examples underscore the importance of uniform implementation of standards.

Being able to share data implies those resources will be easy to locate and readily available when needed. Investigators could make their data available though laboratory-situated computers and generate local backup copies of data. But without a uniform approach for storing, organizing, and serving up the data, such collections will remain isolated and be difficult for others to find or use.

Research data grows at a ferocious pace. Academic Health Centers (AHCs) have been fortunate that the cost of storing these data has fallen faster than storage needs have grown. Enormous volumes of stored data require prodigious networks to serve them up or move them around. Network bandwidth remains an expensive commodity. Despite universally adopted networking standards, data sharing is likely to be limited by institutional investments in network infrastructure, such as very high-speed switches, routers, data cabling, and bandwidth procured from the commercial Internet, Internet2, and National Lambda Rail.

The last section of this report describes how technologies make it possible to securely share information between parties or organizations. Encryption offers researchers a means to safeguard sensitive patient data as it moves from one location to another. Virtual private networks using the standards of Internet Protocol Security, IPSec, or Secure Sockets Layer, SSL, VPN communication links have become standard security practices. Additional means are available to encrypt data stored on computer disks, DVDs, memory sticks, and mobile devices. While encryption standards exist, they may not operate seamlessly between information systems.

It is a simple matter to store digital data for a few years by writing them to magnetic or optical media. The library community can store and retrieve paper documents for 100
years or more. Archiving data for decades or generations on the same digital medium remains an elusive goal. Information written to magnet media must be refreshed or rewritten periodically to reduce the risk of loss. Optical data storage media may need fewer frequent refresh cycles. Even though data bits and bytes remain intact, formats used to read or interpret these data may have disappeared long ago. Application vendors go out of business. Obsolete data formats fall by the wayside. Selecting technologies from open source and commercial vendors that have demonstrated a history of faithfully adhering to standards and preserving compatibility (e.g., TIFF, HTML, XML, and PDF) is an important data preservation strategy.

Assembling collections of data backups into a makeshift archive can build a satisfactory research data archive adequate for reuse in that laboratory. This approach is inefficient and does scale across an organization. Maintaining an institutional research data archive is important not only for sharing research observations but also for establishing intellectual property claims and demonstrating research compliance. A properly documented and maintained archive offers proof of discovery and priority. U.S. government agencies typically require data to be preserved for some specified time after completion of a project. Research data retention and archiving policies are necessary to ensure that specific contractual or regulatory requirements will be met. A decision to consider is whether to archive raw experimental along with the final results of analysis. A strong case can be made for preserving raw data because access to original observations enables scientists to compare analytical methods and possibly derive new insights without repeating costly experiments.

Recipients of research data sets must be able to interpret correctly the information they contain. Information systems may store or manipulate identical data using slightly different interpretations of the same standard. As mentioned previously, organizations find it difficult to exchange all but the most basic elements of patient information between different systems using the same HL-7 standard. Initiatives by the Clinical Data Interchange Standards Consortium (CDISC), the Certification Commission for Healthcare Information Technology (CCHIT), and the Health Information Technology Standards Panel (HITSP) seek to harmonize standards and their interpretation to make interchange of clinical and research data simple and more reliable. These and similar efforts that seek to bring uniformity among implementations of standards should be encouraged.

**Technology Guidance**

- Provide adequate data storage and network bandwidth to accommodate expected growth of research collections and their movement between research partners.
- Distinguish between data backup and archive needs and develop standardized solutions for each.
- Employ widely adopted standards and implement them harmoniously with external organizations whenever possible.
- Support further harmonization of standards to improve system interoperability.
- Provide or arrange for others to provide central data hosting and archiving services to the local research community.

**Processes**

Research sponsors mandate data sharing. Individual investigators using standardized tools in agreed-upon ways to collect research data locally is an adequate beginning. However, when researchers apply ad hoc methods to organize, represent, or manipulate their data, results can become uninterpretable even to researchers in the same discipline. The lack of uniform end-to-end standardization around processes that transform raw data into useful information inhibits sharing and reuse of experimental results.

Assembling collections of data backups into a makeshift archive can build a satisfactory research data archive adequate for reuse in that laboratory. This approach is inefficient and does scale across an organization. Maintaining an institutional research data archive is important not only for sharing research observations but also for establishing intellectual property claims and demonstrating research compliance. A properly documented and maintained archive offers proof of discovery and priority. U.S. government agencies typically require data to be preserved for some specified time after completion of a project. Research data retention and archiving policies are necessary to ensure that specific contractual or regulatory requirements will be met. A decision to consider is whether to archive raw experimental along with the final results of analysis. A strong case can be made for preserving raw data because access to original observations enables scientists to compare analytical methods and possibly derive new insights without repeating costly experiments.

Recipients of research data sets must be able to interpret correctly the information they contain. Information systems may store or manipulate identical data using slightly different interpretations of the same standard. As mentioned previously, organizations find it difficult to exchange all but the most basic elements of patient information between different systems using the same HL-7 standard. Initiatives by the Clinical Data Interchange Standards Consortium (CDISC), the Certification Commission for Healthcare Information Technology (CCHIT), and the Health Information Technology Standards Panel (HITSP) seek to harmonize standards and their interpretation to make interchange of clinical and research data simple and more reliable. These and similar efforts that seek to bring uniformity among implementations of standards should be encouraged.
Data and system compatibility issues such as these led research consortia such as the Cancer Biomedical Informatics Grid (caBIG) and the Biomedical Informatics Research Network (BIRN) to seek better ways to represent and exchange many types of information among collaborating institutions. Domain-specific ontologies together with controlled vocabularies such as LOINC and SNOMED enable them to achieve compatibility of data sets across organizational boundaries. Including computer-usable definitions of basic domain-specific terms and relationships further promotes the unambiguous interpretation of information. Organizations such as the National Center for Biomedical Ontology (NCBiO) offer comprehensive information on ontologies.

Metadata are artifacts that describe an associated data set. Metadata may detail the experimental conditions under which data were obtained. They are a necessary component of data sharing because subtle differences in materials or methods lead to conflicting data interpretations. Like ontologies and controlled vocabularies, metadata extend the usefulness of shared experimental data.

The fundamental academic goals of discovering and disseminating new knowledge conflict with the goals of preserving confidentiality and protecting intellectual property. The decision of what data to share and with whom to share it must respect applicable laws, regulations, contractual obligations, and organizational policies. Policies must also consider the perceived value of the data, the economic interests of the parties involved, and the benefits to society. Reconciling these competing goals requires reaching consensus among all parties involved.

At the birth of collaborative research, it was possible to negotiate data sharing and confidentiality agreements between pairs of organizations. As the number of institutions in a research consortium grows, the number of pairwise agreements balloons exponentially. An alternate approach for today’s clinical translational research initiatives can be found in adoption of common policies and shared governance.

Fundamental tension between publicity and secrecy, inclusiveness, and exclusivity must be addressed when writing data-sharing policies. caBIG’s Data Sharing and Intellectual Capital (DSIC) workgroup drafted common data access and sharing guidelines that member organizations can agree to adhere to. For instance, DSIC recommends assessing four characteristics of a data set to determine their “sensitivity.” These include economic, proprietary, and intellectual property value; data privacy, confidentiality, and security; IRB requirements and institutional restrictions; and sponsor restrictions. This is a model that newly forming research consortia could also adopt. Once collaborating organizations agree on a common set of data classification and protection standards, sharing data between many parties can be expedited.

No one office is likely to have the requisite expertise to weight all the factors and craft common guidance. The institutional IRB, HIPAA Privacy Official, Information Security Official, legal counsel, and intellectual property officer should begin to wrestle with these complex issues. Loshin (2002) lists several parties for whom research data has value and urges their interests be considered when formulating policy. Fishbein (1991) encourages institutions to state clearly their policies and offers guidance for creating them.

### Process Guidance

- Develop a policy defining ownership of research data along with the stewardship responsibilities that follow from it.
- Develop policies addressing data sensitivity and its value to parties who create, share, and consume those data.
- Establish measures commensurate with data sensitivity and value to assure the confidentiality, security, integrity, and availability of research data sets.
- Help create frameworks that enable automated authorization to digital resources.
People

People participate in many different capacities during creation of biomedical research data. A laboratory technician or graduate student may prepare samples and perform measurements of discrete values. A biostatistician analyzes these values and tests hypotheses. The research team distills and integrates these data along with those from other sources while preparing a manuscript for publication. Each member of the team plays a significant role. Some members may have more than one role. The list of contributors together with the description of methods creates a pedigree sufficient for other investigators to interpret and rely on those data.

Authorization, the process of determining which authenticated individuals are granted access to which digital resources, depends on an individual's credentials and role(s) within a project. Some roles such as principal investigator may be generic. Others can be highly project-specific information. Describing each individual's role and their relevant professional credentials should accompany an electronic data set whenever it is shared. This information, as cited in the previous section, might be available from many sources and could one day be acquired by automated means.

Automating authentication and establishing data provenance, above, are key next steps but ones where much progress remains to be made. Authorizing levels of access to research data sets remains a manual process today in large part because objective verification of a person’s professional attributes is insufficient to build a desired degree of trust. Establishing trust among ad hoc collaborative teams spanning many organizations is a cultural and legal barrier that must be overcome if collaborative research is to flourish more broadly.

Moving Forward Requires Alignment

Together, technology, processes and people comprise research information systems. Each of these components must serve certain data protection, management and stewardship obligations if collaboration is to function well. No one component can do everything. The three categories must mesh to create a hospitable data sharing environment in which inter-institutional research collaborations flourish.

Local resources and specific project requirements define expectations for technology, process, and people. Sharing research data between organizations raises issues of differing implementations of standards, conflicting policies, multiple roles, and disproportionate economic interests. Until recently, such questions were addressed ad hoc within universities or between organizational partners. As the volume of research data, the number of investigators, and the diversity of organizations rises, ad hoc agreements become unwieldy. There is a need and an opportunity to forge a consensus of common expectations around research data sharing. Some of these are given here as Guidance at the end of each section.

We recommend that organizations build on existing policies and standards extending them as appropriate to address the challenges ahead. Institutions benefitting from harmonizing their actions will prosper from these efforts. Failure to reach a broad consensus will slow data sharing, limit research collaborations, impede discoveries, and delay benefits that society expects.

People Guidance

- Document the credentials of a research team so data provenance can be tracked.
- Work to build trust among collaborating individuals and organizations.
Communication and Collaboration Technologies

Modern biomedical research is increasingly being conducted by highly distributed, and frequently, multinational teams of investigators. Fortunately, the widespread availability of the Internet and high-speed networks, together with the proliferation of online information and data resources, are fostering the use of an ever-increasing array of electronic collaboration tools. The explosion of new electronic collaboration technologies and tools is providing researchers with access to a wide range of options for supporting this new model for research. In particular, these tools are allowing investigators to bridge the constraints of distance and time. But amidst the constant evolution of both the users’ expectations and the products themselves, no one product has emerged to meet all needs. As a result, any current project requires a complete individual assessment of its needs for interoperability, security, scalability, and compliance requirements to identify functional and cost-effective communication and collaboration solutions.

The following sections will briefly examine some of the current and emerging communication and collaboration tools used to support modern biomedical research. This is not meant to be an exhaustive catalogue of products but an overview of the general types of technologies, the needs that they address, and the issues they present as they grow in popularity. Many of the technologies described have become ubiquitously available; but few, if any, of these in their native form will likely meet institutional requirements for security, privacy, or access control. Therefore, it is extremely important that each technology solution is assessed in the context of the particular research team involved – including the people on the team; the types of information being collected and stored; the services required; and the risks associated with the inappropriate access to data, information, results, or the workings of the group itself. For some research team activities, the use of certain publically accessible tools might be perfectly acceptable, requiring only the proper configuration of security and privacy settings. For other research teams, the use of these same tools could represent a major information security risk and a significant liability for their institutions.

Another overarching consideration for all communication and collaboration tools is their ability to utilize federated identity management models. Individually managing identities, particularly for larger teams working across several organizations, is onerous and also presents significant security risks. The ability to successfully use any of these tools and access resources that are being shared by the team will require an active commitment to implementing the needed IT infrastructure and developing the policies and procedures critical for the successful and appropriate use of any of these technologies. Review of this area is subdivided into four main areas: Communication and Information Sharing Tools, Collaboration Tool Suites, Web 2.0/3.0, and Technology Infrastructure.

Communication and Information Sharing Tools

Communication tools have traditionally been divided into two categories based on the immediacy of the interactions that they support. Asynchronous technologies are those that support work among individuals or teams, without requiring interactions to happen at the same moment. These allow investigators to “time-shift” collaborative interactions to better accommodate their individual schedules or compensate for the problems associated with coordinating real-time interactions across multiple time zones. Technologies that support collaborative work among individuals or teams in real-time are termed synchronous. However, distinctions between “asynchronous” and “synchronous” technologies are now being blurred. A traditionally asynchronous technology, such as e-mail, is routinely being used to exchange documents during voice conference calls, virtually in real time. Likewise, video or Web-based conferencing systems often have the ability to simultaneously stream sessions over the Web in near real time or viewed asynchronously “on demand” at a later time. The following sections review some of the more common types of
communication and information-sharing tools being used by research teams.

E-mail and File Sharing

E-mail is probably the most commonly used communication and information-sharing tool for researchers. The popularity of e-mail is due to its almost universal availability, low cost, and convenience. Investigators in developed and developing countries now have access to e-mail through their institutions, commercial Internet, or cellular service providers, or through e-mail services such as Gmail or Yahoo. One of the major deficiencies of e-mail is its inability to securely and efficiently deliver the large attachments (e.g., data files, images, publications) associated with research projects. Many e-mail systems, by design, limit attachment sizes or give a very low transmission priority to messages with large attachments. In addition, most e-mail solutions do not protect sensitive information during end-to-end transmission or ensure the authenticity of the source. To address these shortcomings, a number of different solutions for file transfer are available and are being implemented by institutions and by individual research teams as a complement to their e-mail systems. These include external Web-based services (such as Xythos BearSpace), hardware appliances linked to an institution’s e-mail system (such as Accellion), and separate software products (such as the Cyberfusion Integration server). Augmenting e-mail communications, instant messaging, chat, and text messaging are readily available and used by many research teams.

Voice and Video Conferencing

The convergence of data and voice technologies and the rapid growth of cellular coverage are placing modern collaboration technologies in the hands of researchers around the globe. In 2010, the number of “active phone lines” is expected to be 7 billion, equal to one line for each person on the planet and, currently, over 1 billion cellular subscribers have access to data-enabled cellular networks. Conference calling using standard telephone technologies or Internet-based calling has become the mainstay for research collaboration. International calling rates from many countries are dropping in price, and the widespread availability of Internet-based calling services such as Skype, which also allows call-out to standard telephone lines, gives international teams of investigators ready access to conference calling capabilities.

Although voice conferencing appears to meet most of the communication needs of many research teams, the use of video conferencing technologies is often a preferred means of synchronous communication. Special-purpose video conferencing systems, using Internet-based protocols, are now capable of delivering very high-definition video with outstanding sound quality, but these systems are expensive and are typically available only within institutional videoconferencing facilities. The convenience of engaging in virtual team meetings, from your desk or while traveling, has prompted the growth of a desktop-based or mobile phone-based video conferencing market. Low-cost computer video cameras and head sets can turn almost any computer into a videoconferencing system. In addition, many cellular phones now come equipped with video cameras, making mobile videoconferencing eminently feasible. Although there are trade-offs in quality, the convenience and accessibility has made these lower-cost systems very popular. The ability of Skype and the Mac-centric ICChat protocol to support low- or no-cost-videoconferencing and the incorporation of 802.11 WiFi into cellular phones makes these the methods of choice for many research teams. A major challenge with many of these tools is that IT departments must install clients on institutional systems and/or open ports for accessibility, which might conflict with institutional security strategies and policies. Many of these products also need special configurations to reduce loads on the networks or improve security.

The need for research teams to jointly visualize complex data sets or conceptualize system interactions has been one of the reasons for the development of more advanced videoconferencing technologies. Virtual telepresence, shared visualization walls, and digital video over IP (DVTS), often supported by special regionally based networking services, have emerged as enhanced alternatives to standard
video conferencing technologies. Expensive infrastructure requirements and high maintenance costs are restricting the rate of adoption of these technologies, but lower-cost or shared desktop solutions are being developed.

Web Conferencing

Web conferencing technologies are now playing a major role in facilitating the work of biomedical research teams. Products and services such as Cisco’s WebEx Meeting Center, Adobe Acrobat Connect, and Citrix’s GoToMeeting support one-to-one as well as one-to-many conferencing scenarios. These integrated voice-over IP or dial-in conferences allow for team meetings to be held at any time with team members located at almost any location around the globe for very reasonable costs. In addition to having multisite video conferencing at their desks, individual team members can also fully share documents, applications, presentations, and electronic whiteboard activity with the group. Electronic whiteboards or tablets, such as those produced by Panasonic, Smarttech, Mimio, and Hitachi, are interactive displays connected to a computer. These devices are becoming increasingly popular for facilitating team activities. By touching the board or writing on the display, information is shared with others, often at remote locations. Remote users may also interact with their displays to share information.

Collaboration and Project Workflow Tools

Authoring Tools

Commonly used document-authoring software, such as Microsoft Office, Adobe Acrobat, and Google Docs, have many features that facilitate collaborative authoring of individual documents. These include the ability to track changes by author, add individual notes to the document, work in different languages, and utilize document templates. The ability of each of these tools to provide appropriate levels of security and to effectively track and manage multiple documents, images, videos, large data sets, and other types of information associated with a complex research project is generally limited, depending on the particular product and its configuration.

Learning Management Systems (LMS), which are used by many academic centers to support educational programs, are frequently repurposed by research groups to support their collaborative interactions. These systems generally allow for collaborative authoring, with check-in/out of content objects, and also provide discussion lists, with threaded discussions. Tools such as discussion lists have been found to be a good way to identify, resolve, and track disagreements as they arise in the authoring process. Since many faculty are already trained on these systems and the user interfaces are designed to be particularly intuitive, most find the threshold for adoption to be low. Research data privacy requirements and licensing between organizations are often the more challenging issues facing the repurposing of an LMS for research needs. Encouragingly, many LMS products are now incorporating a much wider array of tools that can be used for research collaboration management, including metadata for individual objects, links to the literature and other information resources, search features to rapidly locate needed information, regular archiving of information, and audit trails for compliance. Particularly useful for supporting cross-institutional collaborations is the ability of many LMS products to utilize federated identity management services (addressed in the next section). Although not specifically designed to support collaborative authoring and the work of research teams, LMS can provide a low-threshold interim solution at many institutions.
More recently, Web page creation and management systems, such as Wikis, and blogs have become popular for collaborative authoring and workflow tracking. A Wiki is a collection of Web pages designed to enable team members to contribute or modify content via a user-friendly interface. Wiki services, such as Confluence or MediaWiki, are easily accessible via the Internet and are also run locally by many institutions. Wikis and other tools that support Web-based communities are being bundled with e-mail services, online document authoring tools, shared directories and calendars, and survey functionality to provide ad hoc collaboration tool suite functionality useful to support the activities of research teams.

Collaboration Suites

The coordination of research team activities and the management and assembly of the vast amounts of information associated with a research project are often accomplished using specific collaboration and workflow management software suites. Collaboration tool suites, also called groupware, have been widely used in business since the 1980s. Tool suites, such as IBM Lotus Quickr, EMC’s Documentum Centerstage, and Microsoft Windows SharePoint Services, have functioned well within organizational boundaries but have not been widely shared across institutions due to licensing issues. Atlassian’s Confluence Wiki linked with their Jira issue tracking and project management software is becoming a popular combination for supporting team research at academic institutions. Other Web-based collaboration tool suites include Community Zero, Novell’s SiteScape, Jive’s Clearspace, Documentum’s eRooms, Tomoye, and Zimbra 5.0.

Increasingly, the leading collaboration software suite vendors are embedding content management and records management functionality within their suites to address regulatory compliance. This functionality provides the retention and destruction of records based on a schedule, controlled access to information, on-demand information discovery, and auditing. These products are also incorporating more social networking functionality, but challenges still remain in integrating policy and federated identity management solutions into these systems to provide support for diverse and multi-institutional research teams.

Research Workflow Management

The growing complexity and regulatory requirements associated with multinational, collaborative research projects, particularly those involving protected or sensitive information, requires the linking of project and process management technologies, often across institutions. Common collaboration management tools, such as shared calendaring, contact list management, task assignments, milestone tracking, and dashboards need to function across institutional boundaries.

In addition, the use of common or linked laboratory information and financial management systems, electronic IRB systems, and other services will be needed to effectively coordinate research activities. Regulatory and intellectual property constraints on the use and sharing of patient and research data need to be considered in any collaborative research project. These constraints differ by institution and by country and can be very complex. Process modeling tools as well as attention to workflow integration will be required to assure that electronic collaboration technologies are used appropriately.

Modern biomedical and clinical research is becoming increasingly data intensive, requiring extensive data searching, analysis, and simulation. These processes become even more challenging when they need to be done collaboratively across institutions. Data analysis, search, and simulation must be accomplished using data repositories or warehouses located at multiple institutions. These processes require standardized ontologies, ways for managing and tracking access, and policies for data sharing. The large national biomedical research initiatives, such as National Center for Research Resource’s Biomedical Informatics Research Network (BIRN), The National Cancer Institute’s caBIG, and the NIH’s Genome Wide Association Study, have established infrastructure frameworks and standards to facilitate information-sharing and utilization among members of distinct research communities. Projects such as i2b2 are promising to provide...
similar access tools for clinical and translational sciences research teams dealing with the much more challenging issues associated with patient information.

As information discovery and data analysis processes have become more complicated, researchers have developed workflows and scripts that can automate routine activities. The captured/computed scientific data and methods can be shared with others. In fact, the specific workflow templates used in an experiment are now becoming important experimental parameters themselves, which need to be developed collaboratively, tracked, and assessed. In this context, Taverna, part of the myGrid project, and other workflow systems, such as Kepler, Triana, ActiveBPEL, are applications used to build and execute workflows accessible to the research community. They allow for the creation, execution, and reusability of workflows by combining several services in a coordinated, well-defined manner. Publishing a workflow enables work to be reproduced or at least correctly interpreted and is becoming increasingly important for modern biomedical research.

Social Networking, Virtual Reality, and the Semantic Web (Web 2.0/3.0)

Emerging as potential collaboration tools for the biomedical research community are the increasingly popular social networking and virtual world applications. Widely available Web 2.0 social networking services such as Facebook, LinkedIn, and MySpace are being used by communities of researchers but are not well-suited to meet the needs of team-based collaborative research. Many purpose-built networking sites are now available for scientists, and ones like myExperiment even address the special requirements of the scientists who need to describe the attribution of work, control visibility, and sharing in groups; handle licensing; and work with distributed collections of data. This site also enables investigators to exchange scripts and workflows, allowing others to provide recommendations and maintain the workflows and Web services from various sources.

Slowly being adopted by collaborative research teams are the virtual reality (VR) environments, such as Second Life. Second Life is now being used by researchers to locate jobs, find collaborators, share resources, or interact as virtual teams to address complex problems. The use of Second Life is convenient and inexpensive and has provided a way for scientists to interact with community members around a research topic (e.g., MICA Website: http://www.mica-ww.org/). But this technology also promises to be particularly useful in providing opportunities for scientific visualization. As research data sets become more complex, high-resolution visualization technologies are being used to identify novel patterns, relationships, and interactions. Investigators can also jointly run simulations and even experience multidimensional data sets in the virtual world [Djorgovski, et al., 2009]. Advances in the development of technologies to support immersive VR environments will undoubtedly lead to higher resolution and more realistic environments, which should help propel adoption of this technology as a useful tool for scientific research.

On the near horizon for Web 2.0 collaboration software is the much-touted Google Wave. This planned open source

Collaboration and Project Workflow Tools Guidance

- Set up a properly secured and managed Wiki or collaboration tool suite within your organization, preferably one that can federate with similar collaboration tools at other organizations.
- Follow the activities of the major grid initiatives and the development of standards for the management and sharing of various types of research data (bioinformatics, clinical research, patient care, etc.). Where feasible, try to ensure that internal data stores will be compatible.
- Provide researchers with training and guidance on the appropriate use of electronic collaboration and workflow management tools to support team-based research activities.
product is designed to provide user-defined collaborators with concurrent rich-text editing, where you see on your screen nearly instantly what your fellow collaborators are adding to your wave. In addition, all interactions are recorded for playback providing both communication and collaboration capabilities. A key feature of the Google Wave protocol is that it is designed for open federation, such that anyone's Wave services can interoperate with each other and with the Google Wave service.

Semantic Web and Virtual Worlds are emerging Web technologies, components of what is being called Web 3.0, which hold great promise for facilitating team-based scientific discovery and collaborative research. The goal of the Semantic Web is to define the meaning (semantics) of information and services on the Web, making it possible for people and, more importantly, machines, to analyze and capture relevant information items, identifying the links between individual items and the people interested in them. Such technologies are already starting to appear in the social networking space, with Semantic Web-based services such as TWINE. Realizing the importance of the Semantic Web for health care and research, the W3C, which is overseeing the development of the Semantic Web, has formed a Semantic Web Health Care and Life Sciences (HCLS) Interest Group (http://www.w3.org/2001/sw/hcls/). One of the major benefits of Semantic Web-based databases and application is that they enable machine learning and the development of “intelligent agents.” “Learning” from your interactions with various information objects and people, these agents can search for other linkages to provide researchers with a more complete set of applicable information resources. This will be particularly important for multidisciplinary research where novel connections between experimental disciplines, technologies, and researchers are difficult to identify.

**Social Networking, Virtual Reality, and the Semantic Web Tools Guidance**

- Be informed about the research communities being used in Second Life and other virtual worlds that are relevant to your institution’s research programs.
- Keep abreast of emerging Web 2.0 and 3.0 technologies and how they are being used to promote the work of biomedical research teams.

**IT Infrastructure for Team Research**

**Research Portals (Intranet and Extranet)**

Dedicated research portals are becoming an increasingly important collaboration tool for investigators. These are being established to provide services to individuals within individual institutional networks (Intranet) or more broadly to external communities (Extranet). Products such as Akiva’s WebBoard are striving to meet the needs of both internal and external collaborations, however, most institutions are still using different tools for their internal and external users. These portals enable users to:

- Find new funding opportunities
- Search for potential collaborators
- Locate research equipment, core facilities, and other resources
- Construct and host investigator and project Web sites
- Link to research administrative applications, such as Institutional Review Board (IRB), Institutional Animal Care and Use Committee (IACUC), Clinical Trials, Conflict of Interest, and Environmental Health and Safety
- Link to a larger external community to research activities
- Access shared open-source or licensed software
- Access distributed data grids such as CaGRID.
Challenges and Opportunities for New Collaborative Science Models

Storage and Networking

A decade ago, a laboratory instrument might produce several megabytes of information from one experiment. A confocal microscope generates 10-100 times more information. Today's modern gene sequencers assemble more than 1 terabyte of data from a single sample. Explosive growth in the volume of research data is common at the lab bench and is likely to be repeated soon at the patient bed. Raw experimental data must be stored, indexed, archived, processed, and analyzed. At each step, additional new information is generated, which itself must be managed. In many situations, data growth outstrips the capacity and ability of investigators to store and manage this information effectively.

Research projects require tracking of data from their collection through final publication and beyond. As research data grows beyond investigators' capacity to manage it themselves, the biomedical research community looks to central information technology organizations for help.

During the 20th century, sharing research data with the scholarly community meant publishing summaries of scientific work in paper-based journals. Space, color, and cost considerations prohibited inclusion of much useful data. Statistical summaries and representative examples were all that readers could expect. The rapid rise of silicon-based technologies means it is possible to share vast amounts of experimental data as soon as work is completed. Information not included in an e-journal article can be downloaded from an online repository. Computing and networking technologies make it possible to bring together data from many sources to answer complex questions about health and disease.

Maintaining adequate storage for vast quantities of research data is a matter of scale. It is common to find collections of network attached storage (NAS) and storage area network (SAN) devices in individual laboratories holding terabytes of research data. Central research IT organizations routinely house data sets of more than 100 TB. InteractiveData Corp., an information technology market research firm, forecasts a 61 percent compound annual growth rate for data. [“IDC’s Enterprise Disk Storage Consumption Model: Analytics and Content Deposits Provide a New Perspective on the Future of Storage Solutions,” Doc # 214066, Oct. 2008] If data storage needs continue to grow at this rate each year, petabyte-size equipment will soon be necessary. Kryder's Law, analogous to Moore's Law for microprocessors, predicts that hard-disk capacity doubles every year. The good news is that declining cost of hard-drive capacity makes it easy for investigators or their institutions to add more storage capacity. The bad news is that reading and writing data to disk is constrained by Moore’s Law, bandwidth between computers and hard drives, and by the rotational speed of hard-drive platters. This may make storing vast amounts of research feasible, but using massive data sets challenging.

Moving massive data sets within and between organizations can be problematic. Copying a 1 GB file between two servers takes just a few seconds over a fast LAN connection. Moving a 1 TB data set across that same LAN may take hours. Bandwidth can be a barrier to certain research activities. When bandwidth is the rate-limiting factor, exchanging data media (magnetic tape, external hard drives) may be more cost-effective than LAN/WAN transfers. As media and the devices that read/write it become obsolete, investigators and organizations should plan to migrate data to newer media and devices to preserve compatibility.

Very large data sets present other challenges. Indexing, searching, formatting, and reporting becomes technically challenging. Retrieving an archival copy of original data from a project may be as straightforward as downloading one large file from a Web site. Using database software to add structure and facilitate the location of particular pieces of information in a multi-TB data repository is quite common. Additionally, a variety of databases (PostgreSQL, MySQL, MS Access, MS SQL, Oracle, and others) are found in a single organization. Ensuring that research data stored in various databases and formats can be retrieved and integrated requires continual coordination.

Training and Support

Finally, but by no means of lesser importance, is the need to provide sufficient training and support to allow cross-
institutional and multidisciplinary research teams to be successful. It is understood that training in the operation of specific pieces of equipment or software is essential for investigators to realize the full benefits of technology. What is often overlooked is the training in how to effectively manage a virtual team and in ways to facilitate productive team interactions. [Fiore, 2009]. The challenge is in coordinating this training for the entire team, which will increasingly involve multiple institutions as well as investigators from different disciplines with very different levels of experience with technology and forms of social intelligence. Effective leadership, communication, team-building activities, trust, and conflict management are required for a successful team and are even more important for teams interacting primarily with communication and collaboration technologies.

**Infrastructure for Research Guidance**

- Develop an organizational strategy to match storage capability to the rate of growth of research data (clinical and basic science).
- Assess the feasibility of cross-institutional sharing of any research data repositories using applicable national standards and semantic Web ontologies.
- Develop sustainable funding models to support the continued growth of data storage and data storage support services.
- Invest in training investigators and staff on how to work in virtual teams.

**Maturing Market and Managing Risk**

To satisfy the needs of most businesses, collaboration technologies are evolving rapidly. The plethora of new technologies bundled as vendor-supplied tool sets, integrated into various applications, offered by commercial services or available as open-source applications is creating confusion for investigators and for IT professionals alike. The truth is that no one solution or package meets everyone’s needs, few products provide worldwide access, and there are currently no standards for interoperability or the exchange of information. As a result, investigators are using multiple products, thereby putting pressure on IT personnel who don’t yet have the knowledge base for every technology. This will be a major issue for the near term and will only be resolved with the maturity of this field.

Another challenge is the management of risk associated with using various collaboration tools. Institutions will need to assess each tool with respect to the sensitivity of the information being used with it, educate researchers on their responsibilities for information security, and develop policies and procedures to protect information and networks from inappropriate installation. Many collaboration tools and research networks, such as the Grid associated with the NIH’s caBIG project, require the installation of software clients on the desktops and workstations of individual investigators. Frequently, this presents an element of risk, since these clients have the potential for interfering with other network-based services or business applications. Compliance management, including auditing and archiving of collaborative interactions, will likely be expected.

We recommend the AAMC work with AHCs and health information security organizations to foster the development of security risk assessments and “best practice” configurations for rapidly developing communication and collaboration technologies to support team-based science. By developing a process that allows individuals at AHCs to share information about requirements, product experiences, issues, and policies associated with the use of computer-mediated collaboration tools for research, the AAMC may encourage the development of interoperability among the various collaboration tool suites and workflow management software solutions. This, in turn, could drive standardization of formats and data exchange capabilities by the producers of collaboration tools.

To support academic medicine’s desire to move research into a collaborative environment, we also recommend the AAMC provide regular information and analysis on the trends in the use of Web 2.0 and 3.0 technologies in support of the AHC’s
missions and work to develop and disseminate information on effectively managing multidisciplinary and largely virtual research teams.

**Cross-institutional Authentication and Authorization**

Collaborative tools and “shared” information resources require user authentication and authorization for access. However, since many collaborators needing to use these restricted resources are not affiliated with the hosting institution, the question arises as to how external collaborators will be authenticated and authorized to use these resources. This is a complex problem.

Traditionally, individuals do not have digital identities that can be recognized, evaluated, and trusted by information systems other than at their home organization. This results in a variety of chaotic, identity management work-a-rounds that are extremely frustrating for potential users, are grossly inefficient to manage, and are often very insecure.

As research – both clinical and basic science – grows more complex, it is becoming imperative to have systems that allow for the integration of research information between scientific domains. Often data are captured in various laboratories and discrete information systems. Effective support of research, as well as clinical care, requires a substantial and well-designed infrastructure. Much effort often goes into the selection or development of applications specific to particular laboratory modalities, clinical needs, or scientific questions. However, few resources go into the definition and development of enabling infrastructure that will scale in a uniform fashion to handle the data management and integration needs of multiple laboratory modalities.

There are a variety of business and scientific reasons to consider use of common, well-defined legal and technical architecture strategies for enabling communication and data sharing within and between academic health centers. Current models of authentication and authorization have been developed for narrowly defined and specific purposes. Almost exclusively, these areas of focus have been concerned with the security needs internal to an organization or corporate structure. While these mechanisms have been moderately effective within the domains in which they have been developed, to date, a comprehensive consideration of the academic medical community has not been done. Such an effort involves recognition and consideration of the scientific and clinical workflows that routinely occur between institutions, and the development of structures – technical, legal, and procedural – to support these exchanges in a repeatable and scalable fashion.

Development of a multi-institutional common model of authentication and authorization structures would allow teams of clinicians and scientists working either between or within institutions to easily and effectively exchange data and research results in a secure and well-controlled fashion. Such a model could enable the development of workflows that cross institutional boundaries particularly for team-based science initiatives such as those emerging from the CTSA community. A well-defined common model agreed upon by the academic medical community and congruent with similar initiatives in broader academia and government will facilitate the use of national and even international shared resources, such as the TeraGrid, on a local, state, and national scale. Taken together, these factors would facilitate clinical and basic science research by providing an agreed-upon infrastructure for securing clinical data exchanges with the research community.

Infrastructures will require a common trust fabric; common naming conventions for institutions involved; common definitions of roles, titles, and practice credentials (such as MD, RN, LPN, CCRP, etc.); and the legal, policy, financial, and governance structures to support the effort.

**Identity in Cyberspace**

Currently, no generally accepted identity management infrastructure exists in academic medicine to support large-scale, cross-intuitional collaboration in which various aspects of each collaborator’s identity must be known and trusted. This infrastructure is indispensable when authentication,
authorization, digital signatures, information integrity, and individual accountability must occur dynamically across institutional boundaries among many relying parties. Ideally, electronic transactions requiring knowledge of personal identities will be appropriately secure, will protect privacy, and will be as user-friendly as public Web pages.

It is important to distinguish between authenticating a person as a physical identity versus authorizing that person to receive specific privilege(s) based on his or her personal attributes.

1. **Authentication** is a process whereby a relying party can trust at a defined level of assurance (LOA) that a credential truly belongs to the certified physical person presenting that credential.

2. **Authorization** is the process whereby a relying party determines whether an authenticated physical person has the necessary attributes to conduct specific activities using a specific resource.

**Federated Authentication Credentials**

An authentication credential is crucial to establishing a person’s physical identity in cyberspace. Authentication credentials must provide at least the following functions and information:

- Can only be activated by the certified person to whom it was issued
- Positively identifies the credentialing authority (CA)
- Positively identifies the physical person
- Provides a certified unique identifier to the vetted individual that is registered with the CA
- Asserts a defined LOA that the credential is presentable only by the physical person it authenticates.

It is important to realize that a digital authentication credential has far more functionality than an application-specific username/password. A person, or an information system, having a federated authentication credential could use that credential to initiate de novo interactions with relying parties—both people and machines—without further interventions. In the new collaborative milieu within cyberspace, millions of intertwined people and digital systems must safely introduce themselves and interact with others. Widely trusted authentication credentials enable these just-in-time introductions.

**Federated Credential Providers**

Academic institutions must decide how to provide federated authentication credentials to their faculty, staff, students, and affiliates so they can be authenticated elsewhere by relying parties. Because there is not a single, centralized credential authority, institutions are currently faced with the following choices:

- Implement a credential provider (CP) at each institution
- Use an external, possibly commercial, CP to issue authentication credentials
- Employ some combination of the two.

Most institutions today have an internal identity management infrastructure that allows physical identity and personal attribute to be managed. This, in turn, may support single sign-on authentication within an institution so that each user has a single authenticator such as a password or a one-time password device. When used once, the authenticator grants access to several information resources until the user signs off. In most cases, these identity management (IdM) infrastructures were not designed to provide users with universally recognized credentials. However, if these infrastructures adhered to specific federated policies and procedures and were suitably modified, they could be used as authentication credentials between federated institutions. Before relying parties can "trust" locally issued credentials,
Challenges and Opportunities for New Collaborative Science Models

each organization must adhere to well-defined policies and procedures. Such IdM federations comprised of member organizations are forming today.

Systematic use of federated identities is still in its infancy. It is not clear whether it is better for institutions to become a CP themselves and issue their own authentication credentials or to rely instead on external CPs to provide this service for them. Since universities have traditionally handled authentication and authorization activities internally, the trend is to establish their own CPs. One benefit of using an external CP is that individuals can keep the same credential and hence the same unique identifier, as they move from one institution to another.

Authorization across Institutional Boundaries

Service providers (SPs) can implement authorization (i.e., privilege management) in multiple ways. Some SP applications may have an internal role-based system that administrators use to assign roles/privileges to individuals manually. These administrators must determine from one or another trusted source of authority (SoA) whether authenticated persons have the proper attributes to be granted specific access privileges. In order to accomplish this, the administrator uses the person's authentication credential to assure this is the same person about whom the SoA is providing attribute information.

Automating authorization is straightforward when one database holds all the necessary information about every person who might want to use a particular resource. Systems operating strictly within the scope of one organization; for example, employee electronic time sheets, might only need to consult one database, HR/Payroll, to determine whether an authenticated individual needs to track his or her work hours.

A system intended for access by employees at any one of several organizations might need to consult databases at those or other organizations to validate that an authenticated person’s attributes meet stipulated criteria. For example, in order to access a clinical research database, prospective users might have to demonstrate employment in a clinical department at an AAMC-accredited medical school, hold an M.D. degree, be board-certified in pathology, and be the PI on an active NCI grant. Several databases located at independent SoAs would have to be consulted before these criteria could be automatically verified.

SP applications may automate the authorization process by interfacing with attribute authority systems and/or other trusted SoAs. This allows the SP application to query approved SoAs about whether an authenticated individual has the appropriate personal attributes to be granted specific privileges.

A vast number of formal and ad hoc research collaborations will be created in the coming years. If there is any hope for interoperability among them, it will be necessary to establish authentication standards and governance processes. Research systems should be efficient and easy to maintain. Automating authentication and authorization will achieve both.

Authentication Standards

CPs must adhere to clear, well-defined policies and procedures when issuing and managing identity credentials. Organizations can trust credentials from a CP only when it explicitly demonstrates it has adhered to commonly agreed upon standards. Since there is not a “universal” single CP, as is the case in some countries, identity management federations are forming. Within such federations, members agree to adhere to specified open standards and/or openly published frameworks.

In 2003, the Office of Management and Budget published Memorandum M-04-04 entitled “E-Authentication Guidance for Federal Agencies.” This document defines an LOA, 1 through 4, adequate to limit the consequences of authentication errors and misuse of credentials. As the consequences of an authentication error becomes more serious, the required LOA increases. Thus, for example, Level 1 requires no identity vetting whereas Levels 2 thru 4 require progressively stronger vetting of identity, stronger credential binding to the identified individual, as well as progressively stronger authentication tokens.
In 2006, the National Institute of Standards and Technology (NIST) issued NIST Special Publication 800-63 version 1.02 Electronic Authentication Guideline. This document provides specific technical requirements for tokens (typically a cryptographic key or password) for proving identity; identity proofing, registration, and delivery of credentials that bind an identity to a token; remote authentication mechanisms, that is the combination of credential, tokens, and authentication protocols used to establish that a claimant is in fact the subscriber he or she claims to be; and assertion mechanisms used to communicate the results of a remote authentication to other parties. Subsequently, the U.S. Authentication Identity Federation provided an Authentication Credential Assessment Suite to assess compliance of a Credential Service Provider and their Credential Services to E-Authentication Levels of Assurance.

The Password Credential Assessment Profile evaluates Level 1 and Level 2 authentication credentials whereas the Certificate Credential Assessment Profile evaluates Level 3 and Level 4 credentials. The reason for the two profiles is that assurance Levels 1 and 2 apply to assertion-based credentials that use password tokens, (e.g., Shibboleth SAML assertions), whereas Levels 3 and 4 require cryptographic tokens.

Academic medicine is moving toward adopting these federal standards. Shibboleth is open source software that emerged from the NSF Cyberinfrastructure project and has been developed by the Internet2 Middleware Initiative. The development of the Shibboleth Authentication and Authorization infrastructure by the Internet2 makes the implementation of Level 1 and Level 2 authentication credentials relatively easy. InCommon (http://www.incommonfederation.org) is the first nationwide federation for higher education. It utilizes the Shibboleth infrastructure and it is attempting to adhere to Level 1 and Level 2 assurance levels. The University of Texas System Identity Management Federation is a successful example of multiple organizations taking a common automated authentication approach.

In 2007, the National Institutes of Health announced a Memorandum of Agreement (MOA) for inter-federation with InCommon. This announcement noted NIH’s “goal is for researchers to use their institutional identity credentials to authenticate to NIH online applications and services. All NIH online application have been assessed and assigned one of four federal LOA identity requirements for authentication.” (http://www.nih.gov/news/pr/aug2007/cit-14.htm) Subsequently, both the NIH caBIG and the Clinical Translation Science Award (CTSA) Consortium have announced plans to comply with the same requirements.

The AHC community should consider establishing a Shibboleth Identity Provider (IdP) to assert Level 1 and Level 2 authentication credentials for at least some of their workforce. In addition, they should consider enabling various information resources as Shibboleth-enabled SPs that can accept Level 1 or Level 2 authentication credentials. It should also be noted that a number of institutions are successfully requiring providers of commercial software applications to Shibboleth-enable their products.

The most difficult task can be implementing an institutional IdP that meets the federal assurance standards for Levels 1 and 2. Joining an existing Identity Management Federation whose members are collectively working to meet the assurance requirements is likely to be easier than doing it alone. As with any auditable process, it will be crucial for AHCs to document all security-related policies and procedures. Creating or adopting a complete set of Standard Operating Procedures is a critical step in this process.

The NIH MOA with InCommon recognizes InCommon’s Bronze Profile as Level 1 LOA. InCommon’s Sliver Profile is designed to meet Level 2 requirements, however, not all policies and procedures have been agreed upon for the membership. Institutions that already have a solid institutional IdM system in place should be able to easily implement and integrate Shibboleth into that infrastructure.

Institutions should consider using a commercial Shibboleth credential service provider if local resources do not permit a locally hosted IdM infrastructure to credential individuals who are not institutional personnel. Some IdM federations are considering accepting commercial IdPs as federation members in recognition of these needs.
Preparing the Research Environment for Collaboration

Prepare for cross-institutional workflow and transactional data exchanges

Necessary steps to prepare for this process include developing a strategy for IdM that recognizes the growing importance of federations and extra-institutional sources of authentication and authorization. This strategy should closely examine the U.S. Federal Government e-Authentication Initiative and NIST standards. A central consideration is the use of digital signatures, which require digital credentials based on a public key infrastructure and some form of cryptographic key management systems. These measures will likely become necessary for maintaining the security and privacy of medical records and other sensitive documents. This will evolve from traditional username/password approaches commonly in use.

Build directory-aware systems

From a technical roadmap standpoint, consider developing a list of core business and scientific applications that are “aware” of directory services (such as LDAP or Active-Directory). Develop a single-point authentication service for, at minimum, all Web-based applications. Newly acquired applications must be directory-services aware, and this should be treated as a high-priority functional requirement. Examine existing applications to determine which are the important SoAs for key information involved in authentication. Many campuses have found that often there is more than one system involved in such decisions for different groups of individuals. Developing a list of these key systems and documenting the processes involved in provisioning them with this information will allow them to be both better controlled and to be interfaced to control and provision the directory services in a more efficient manner.

Become familiar with Shibboleth

Nearly all emerging federations, including the InCommon Federation, the NIH InCommon Federation, the United Kingdom’s JISC, NCI’s caBIG™ effort, and a number of European grid computing projects have Shibboleth at their core. A number of open source collaboration platforms are well integrated with this software and are available at the Internet2 Web site. A growing number of enterprise and other software systems are becoming Shibboleth-aware as well, as the 116 InCommon-member campuses continue to promote adoption of this technology.

Similarly, the underlying protocols for security are based on the use of Security Assertion Markup Language (SAML) version 2.0. Institutions should advise system selection teams to use this as criteria when specifying and evaluating technology.

Adopt standard forms for personal attributes

Examples of attributes include employee status, date of hire, functional role, or any other piece of data deemed important in inter- or intra-institutional authentication or authorization decisions. These and other attributes have been agreed on by InCommon and other organizations. They are defined in the InetOrgPerson and EduPerson schemas.

Authentication and Authorization Guidance

- Focus on external rather than internal identity management
- Develop standard forms for physical identity and personal attributes
- Define authoritative third-party SOAs for trusted information
- Take an active part in federations and participate in building communities of trust
- Achieve and maintain a level of identity assurance commensurate with the security required for the envisioned uses of research data and information systems.
Work Still Needed

Incorporation and governance

The AHC community should consider integrating with the Internet2 InCommon effort. As currently written, InCommon participation agreements are not completely “friendly” to all AAMC-member organizations, particularly those representing nonprofit research organizations not directly part of existing universities or colleges. InCommon’s board is aware of this barrier, but given NIH’s widespread adoption of the InCommon frameworks and with urging from more diverse organizations such as AAMC, this issue could be resolved. Alternatively, institutions may establish governing and participation structures unique to the AAMC and federate with InCommon and the NIH.

A related issue involves how individuals with legitimate need could gain access to a federation’s resources if they are not already affiliated with one of its members. Some individuals may have valid identity credentials from government or other sources acceptable to the federation. InCommon, SAFE-BioPharma Association®, and others already have agreements and infrastructures in place to allow this. The challenge is issuing identity credentials to individuals unaffiliated with a federation member institution. One approach may be to contract with an outside provider of identity credentials.

Identification of missing technical and infrastructure components required to support academic medicine

Current authentication and authorization efforts, whether conducted locally or between institutions, have yet to develop in all facets of academic medicine. In clinical contexts, authorization decisions will require a computerized application to consult multiple SOAs. Some of those SOAs do not yet exist or have no standardized form for such information. While the AAMC cannot fully address this issue, it is in a position to work with other professional and medical organizations to promote development of this infrastructure.

Other areas that require effort will be the development of a variety of taxonomies and ontologies, including:

- Common forms for institutional names, including their constituent parts (work is underway at the NIH to deal with this issue).
- A full spectrum of high-level use cases, analysis of common abstractions, and development of specific role definitions (or attributes) needed for non-clinical uses. Note that HL7 (http://www.hl7.org) and the ASTM have already done substantive work on this topic in the clinical area, so that this effort would not need to duplicate that effort.
- Object class definitions incorporating the roles and attributes suitable for use by directory services and inter-federations. An example is the eduPerson object class development by Internet2 and EDUCAUSE.

We recommend the AAMC consider working on behalf of its member organizations to develop the technical and policy infrastructure to facilitate integration and harmonization with the NIH Interfederation and Internet2 efforts for NIST authentication LOA 1 and 2. Additional efforts might include a blanket agreement for member organizations to become part of the SAFE BioPharma framework, the Liberty Alliance, or an equivalent AAMC-specific federation.
References


Research Task Force Members

William A. Weems, Ph.D.—Chair
Assistant Vice President, Academic Technology
University of Texas Health Science Center at Houston
Associate Dean, Medical School Information Technology

Kenny Chu
Senior Director, IT Security
Mount Sinai Medical Center

David A. Damassa, Ph.D.
Dean, Information Technology
Tufts University School of Medicine

David A. Fenstermacher
Executive Director, Research Informatics
H. Lee Moffitt Cancer Center

John J. Hutton, M.D.
Professor Pediatric and Internal Medicine
Children’s Medical Center-Cincinnati

Juany Jardines
Manager Research Computing
Memorial Sloan Kettering Cancer Center

Thomas Kieber-Emmons
Professor of Pathology
University of Arkansas College of Medicine

Cheryl F. Lane
Director, IT Research
University of Arkansas College of Medicine

Frank Manion
Director, Research Computing Services
Fox Chase Cancer Center

James McNamee, Ph.D.
Associate Dean and CIO
University of Maryland School of Medicine
Research Task Force Members (Continued)

Marc Overcash
Deputy CIO, Research and Health Sciences
Emory University

Mark Weiner
Associate Director, Research Informatics
University of Pennsylvania School of Medicine

Ex-Officio:

Lynn K. Morgan
Vice President, Information Technology
Associate Dean for Information Resources and Systems
Mount Sinai School of Medicine

AAMC Staff:

Morgan Passiment
Director, Information Resources Outreach