

Spotlight: Institutional Data

Section III, goal 2b, "Make institutional data more widely available for use in collaborative contexts," defines targets to make institutional data readily available for use in collaborative contexts. This appendix provides more in-depth discussion regarding those targets.

The institutional data of the campus is a rich and diverse mix:

- Details about the socioeconomic status of high schools attended by students in the campus's incoming freshman classes.
- Comparisons of actual versus planned expenditures in the budgets of campus departments and units.
- Accession records describing the amphibian, reptile, bird, and mammal specimens in the Museum of Vertebrate Zoology.
- Curricula vitae for faculty members and other instructors who have taught campus courses in Near Eastern Art.
- Remote sensor data detailing the uptake of moisture from fog by Coast Redwood trees.

Making the institutional data of the campus appropriately available within collaborative contexts – such as collaborative workspaces, administrative and scientific workflow tools, and customer relationship management systems – yields compelling advantages to the participants in those contexts, and can often dramatically improve the quality and productivity of their work.

At UC Berkeley, campus institutional data—including data related to applicants, students, alumni, and donor prospects—is currently electronically hosted and stored in over 300 sites across the campus. Adding to the complexity of 300+ sites, are the plethora of offices and individuals involved in institutional data management. Each functional unit that manages these data asserts some level of ownership ...

–Invitation to participants in the Institutional Data Management and Governance (IDMG) Task Force from George Breslauer, Executive Vice Chancellor and Provost, and Nathan Brostrom, Vice Chancellor-Administration, September 27, 2007

Paradoxically, given its value, obtaining programmatic access to institutional data, so that it can be used within collaborative and other contexts, is far more difficult than it needs to be:

- It is often difficult to find data;
- It is frequently a daunting task to try to obtain permission to use it; and
- It is too seldom available in a standard way for use by campus or unit-level programmers.

One way to begin to understand the value of using institutional data when collaborating is to highlight the example of bSpace, the collaboration and learning environment provided to the campus by Educational Technology Services (ETS). By drawing upon multiple institutional data sources, that environment makes it easy for instructors to build course sites in which they, their fellow instructors and instructional assistants, and their students can interact

DRAFT

around the course syllabus, readings, discussions, chats, assignments, and more.

These course sites are readily populated with much of the data that the instructors are likely to need. Instructors are provided with rosters of their enrolled or waitlisted students, which are regularly updated as course registrations change. As students enroll, they become members of each of their course sites automatically, and can start using their sites' tools. Instructors can even choose to view photos of enrolled students, taken when the students received their Cal 1 Cards (photo IDs), so they can practice associating names with faces or better remember a student for whom they are writing a letter of recommendation. In these and a great many other ways, the bSpace environment interfaces with multiple sources of institutional data, making the collaborative experiences it offers much more compelling and seamless for its users.

Recommendations to help make the campus's institutional data more readily available for productive use within collaborative contexts include:

- Identify sources and governance of significant institutional data.

Give institutional-level priority to identifying sources of significant institutional data, and to sorting out who is responsible for its ownership and governance.

This work is already underway, spurred in part by the Institutional Data Management and Governance Initiative (IDMG) (<http://administration.berkeley.edu/idmg/about.htm>) which launched in November 2007. Although that initiative is centered around providing campus leaders with access to strategic data, we remain optimistic that it may also yield benefits beyond that central focus.

- Provide a standard process for requesting access to institutional data, and ensure timely response.

It is often difficult for data users to find useful institutional data. Even when they become aware of the availability of that data, however, it may often not be readily evident how they go about requesting permission to use that data.

As mentioned in Section IV: Findings, when a campus programmer sought to obtain access to the data that is used to generate the contents of a public campus website – the Online Schedule of Classes – she found that three campus units claimed ownership of at least part of that data. She eventually had to persuade one of the data owners to grant her access, who then interceded with the other two data owners.

This needs to change. There should be a uniform process through which any data user can request access to data, and receive a timely response from data managers.

In addition, as described below, data that is unambiguously public should be routinely provided for programmatic access, without requiring that data users specifically request such data.

- Make public data routinely available for programmatic access.

The campus should give weight to the premise that "public data" – data that appear to be unambiguously public and unrestricted, which in many instances has been published for some time on campus web sites – should routinely be made available without restriction, and without requiring any specific request from a data user.

Furthermore, such unambiguously public data should be available in structured formats, useful to campus application programmers. Some data providers currently share such public data only in an unstructured manner – via web pages, PDF or Word documents, or other methods that are intended for reading by humans, but are not readily usable by software. While that is a useful first step, it needs follow-through: by also providing the data in formats that allow it to be used within campus applications and collaborative contexts.

If the parties responsible for governance of straightforward public data do not have the resources to initially provide that data through standard methods such as web services (see below), they can initially offer access to the underlying data via low-cost, low-effort mechanisms. They might choose to do so by providing read-only database views, or even read-only access to file shares where source files containing structured representations of just that public data are stored. In that way, campus application developers who need that data can build interim web services or other standard forms of access on top of that data.

- Standardize on simple, consistent methods for providing access to institutional data.

Rather than providing institutional data via many different architectures and mechanisms, the campus should consider making an early and clear commitment to standardize on a small number of standards-based approaches. This decision can be adjusted over time based on technology directions in industry and campus experiences.

Doing so will maximize scarce resources by allowing programmers to focus their learning and expertise building, by encouraging code reuse, and by simplifying documentation for both data providers and users. For example, when campus data application programmers and other data users know they can always access institutional data in more or less the same way, regardless of its source, this makes it much easier for them to get "up and running," rapidly incorporating that data into new collaborative applications and contexts.

Two related recommendations:

- Provide access via REST or SOAP/WS-* web services, supplemented by direct access to underlying data sources where required.

In industry and the consumer space, access to data is increasingly provided via application programming interfaces (APIs) based on either or both of two complementary web services approaches:

- Web services based on the REST architectural model, suitable for operations that perform straightforward data access and manipulation, where data can be modeled as web-accessible resources.
- Web services based on the W3C's SOAP and WS-* specifications, appropriate for situations where their unique advantages are required; for instance, where end-to-end security services are needed, or where multiple services must be federated to perform transactions.

Campus data providers should also consider standardizing on these two methods for all services that provide data to a broad, general audience. Where performance is a critical consideration, or where certain collaborative tools don't yet have the ability to consume data from web services, providing the additional ability to directly access relational databases or other underlying data sources, bypassing web services interfaces, may be desirable or necessary.

It may also prove possible to offer centralized services that make it easier for data providers to share their data via these types of web services. For instance, a service that allows a provider to share a simple database view as a REST- or SOAP-based web service, based on a number of conventions (see below) and defaults, could shorten the 'time to market' for providing service-based interfaces to institutional data.

- Encourage standardization on common conventions for web services.

When offering web services based on REST or SOAP/WS-*, the campus can encourage taking additional step toward maximizing programmer resources, through ease of learning and code reuse, by encouraging standardization on common conventions for those services.

For instance, for REST-based services, standardization on URL patterns and naming conventions for containers and end-point resources would make it faster and easier for programmers to make use of new data sources. For SOAP-based services, standardization on naming conventions for remote procedures and variable names and data types would offer similar benefits. Coalescing around common conventions for web services that provide access to institutional data can be encouraged by guidelines that emerge from campus communities of practice consisting of data providers and data users, or promoted or required through other means.

- Encourage standardization on data formats.

Where feasible, every web service which provides access to institutional data should provide at least one common data format. For instance, this could be an XML format whose schema is mechanically derived from an underlying database view, or another, similar format.

In addition, each service may also provide additional data formats appropriate to the underlying data or which are most useful to its customer community of data users. For instance, services whose data may often be consumed by feeds might offer data in Atom or RSS feed formats, while services offering calendar event data might offer data in iCalendar format, which can be consumed by many calendaring and scheduling applications.

- Provide incentives for sharing data.

Within institutions of all kinds, not just at UC Berkeley, there is an understandable, even natural, bias against making data available to others. This is because maintaining strict control over the release of data can variously represent:

DRAFT

- Power: if data isn't made available widely, it can be traded to other data providers or users for services or influence.
- Job security: gatekeepers of data have a well-defined niche and value in an organization.
- Appropriate caution: while there are few rewards for making data available, there are many deleterious consequences to inappropriately releasing restricted or sensitive data.
- Protection of resources: making data available, in the absence of a clearly defined need on the part of an influential data user, costs a campus unit both time and money.

If institutional data is to become more widely available for use in collaborative contexts, it may be desirable to consider institutional- and unit-level incentives for providing appropriate access to data. These incentives could be provided through formal recognition during employee performance awards, or Spot Awards and other similar performance bonuses.

Incentives for providing access to data might be based on how many campus applications, collaborative and otherwise, make productive use of that data; or on the strategic value of innovative new applications or enhancements to existing applications that became possible as a result of that data's release and ongoing availability. They may also take into account the quality of the data provided, the extensiveness and quality of documentation around the data, and the ease with which campus programmers can access that data through simple, consistent services.

Data users can be provided with incentives, as well, to provide feedback to data managers and providers about their uses of and satisfaction with data services. Finally, best practices in making data available for productive use can be highlighted and shared within the relevant communities of data managers and providers, and data users.

- Continue work to provide guidance and tools for handling institutional data, and to rationalize its data elements.

The campus has made significant strides toward providing guidance and tools to assist campus data managers in appropriately handling campus data. Campus guidelines specify how data should be stored and managed and what data may be released, to whom, and under what circumstances. There has also been considerable recent activity specifically around identifying and appropriately protecting restricted or sensitive data. Finally, there is ongoing work and guidance related to metadata ("data about the data"), such as identifying where the use of data dictionaries may be necessary, so that the meaning and values of data elements may be correctly understood; and encouraging the standardization – and hence the interoperability – of data elements across multiple data sources.

This work has in part been driven by the campus's Data Integration Initiative in 2002 and the subsequent formation of the Data Stewardship Council, as well as by a number of campus units at the center of the campus's ongoing response to protect restricted and sensitive data, following several highly publicized security breaches which exposed such data. The campus is encouraged to continue to dedicate significant resources to these salutary efforts, all of which will help create a data rich environment in which online collaboration via collaborative tools can thrive.

DRAFT