

**NEW FRONTIERS IN THE DIGITAL LIBRARY: SOCIAL AND
ECOLOGICAL DIVERSITY OF THE AMERICAN WEST**

DRAFT PROGRESS REPORT COVERING THE PERIOD NOVEMBER 1,
2003 TO OCTOBER 31, 2004

RESPECTFULLY PRESENTED TO THE WILLIAM AND FLORA
HEWLETT FOUNDATION

BY

THE CALIFORNIA DIGITAL LIBRARY

DANIEL GREENSTEIN (PRINCIPAL INVESTIGATOR) WITH PETER BRANTLEY, ROBIN
CHANDLER, LAINE FARLEY, ROY TENNANT, AND STEVE TOUB

Value proposition

Digital libraries add value to information by organizing it into coherent, highly differentiated collections that are developed to support particular modes of inquiry and/or user communities. Users access these collections with the current generation of sophisticated search tools, but also through a range of highly specialized services.¹

Google, by contrast, adds value to information by amassing it into vast, largely undifferentiated collections which users examine with increasingly sophisticated search tools.

1. Introduction

The American West project seeks first and foremost to assess the foregoing value proposition through the formal evaluation of users' behaviors and needs, and the related development of collections and tool suites that demonstrate the prospects for and usefulness of high-quality, curated online collections.

The work has proceeded along the following lines that form the organization of this first-year report.

1. Assessing the value proposition against the needs and interests of user communities that take an interest in the development and use of openly accessible online information in an educational setting
2. Developing use cases based on the needs assessment
3. Planning the collection
4. Creating functional and technical specifications for essential tools

Work has leveraged a variety of ongoing efforts at the CDL which are reported here where appropriate. It has also had a catalytic effect, stimulating synergistic activities amongst research libraries nationally, most significantly in the Digital Library Federation's Aquifer initiative.² That initiative adopted the aims of the American West project as its own and promises to substantially supplement both the openly accessible digital library collections that can be made available as well as the tool suites that will enable their curation and presentation to user communities.

Significant achievements in the first year include:

- Building partnership among contributing libraries. Although not discussed at any length in this report, the American West project is forging a closely working community among the partnership and extending that partnership in recent months to include research libraries at Emory, Illinois, Johns Hopkins, and Stanford universities.
- Completing initial user needs assessments for the design and construction of the American West collection and associated tool suites (further needs assessments

¹ Thus a collection of online texts may be accessible with a range of analytical features that support computer-based literary and linguistic analyses; a collection of social science data is surrounded with visualization and statistical processing tools, etc.

² The Aquifer initiative is described at <http://www.diglib.org/aquifer/>. Its collection development path is described in the report on the Aquifer meeting, held on October 25, 2004

will be conducted in year 3 based on hands-on use of the American West (Section 2).

- Developing use cases that guide key investment decisions with regard both to collection content, service functionality, and the provision of appropriate tool suites (Section 3).
- Planning the collection (Section 4):
 - reviewing collection content on offer from partner libraries with a view to determining collection strengths and gaps;
 - assessing possible strategies for filling collection gaps; and
 - reviewing the technical issues involved in capturing and aggregating collection content in a manner that supports the service functionality required for the American West project.
- Functional and technical specification for essential tool suites and partial completion of those tool suites (Section 5).

2. Assessing the value proposition

The American West project will assess three simple and closely related hypotheses,

1. Users in an educational setting need coherently organized collections of information that are built specifically to support particular modes of inquiry and/or their user community.
2. Different users and uses require a multiplicity of curated collections (research physicists require different collections than undergraduate physics students or teachers preparing physics lessons for presentation in a high-school science course).
3. The development of high quality, coherently organized collections that satisfy particular users' needs or modes of inquiry may be enabled by the development of tools that enable libraries and other information organizations that support research, teaching, and learning to cost effectively build such collections.

In the past year, the first two of these hypotheses were tested in a series of 18 interviews conducted with 45 potential users, drawn from educational institutions across California and Colorado.³ Participants included:

- academic librarians;
- public librarians;
- university graduate students;
- university faculty;
- community college faculty; and
- K-12 teachers and media specialists.

³ See "Documenting the American West. User Interviews. Final Report (1 July 2004). Prepared by Alex Wright (alex@agwright.com) with Rosalie Lack, Robin Chandler, Ellen Meltzer, Brenda Bailey-Hainer, Steve Toub, and Roy Tennant available from http://www.cdlib.org/inside/projects/amwest/americanwest_assessmentfindings.pdf

The interviews employed the “friendly dyad” format, with pairs of participants teaming up for 60-90 minute interview sessions conducted by a moderator with one observer. No interview session included more than four participants.

The primary assessment aim was to generate qualitative insights about the value to users of a coherently organized collection bearing on the history and ecology of the American West. In particular, the project sought users’ insight into the potential value, utility, and desirability of coherently organized online collections with a particular theme.

Finding that the concept appealed widely, the interviews explored two areas to identify users’ needs for:

- collection scope and breadth (what would have to be in a collection in order to make it useful and valuable to a particular community?), and
- access tools and features (how should collections be organized? What search, browse, export, and other features are required, and how are these prioritized by users in different communities?)

Assessment undertaken for the American West project leveraged and contributed to a broader assessment program being carried out by the CDL to evaluate the value proposition set out above, albeit with regard to different kinds of collections. The American West assessment focused on the prospects for online collections that selectively integrate extant materials bearing on the history and ecology of the American West. The “Undergraduate Core Collection” study asked the same kind of questions to UC undergraduates with regard to a proposed collection of online science information that would support undergraduates who need to complete an assignment in a strictly limited amount of time.⁴ A further study was conducted with academic, public, and community college librarians and with K-12 teachers and media specialists with regard to “curated” collections that could be built from materials presented through the National Science Digital Library (NSDL), and integrated with local holdings (e.g. online journals and reference databases that academic libraries subscribe to).⁵ The assessment activities were aligned with a common framework to ensure comparability of results and to enlarge the number of respondents from whom information was gathered, thereby increasing confidence in our findings.⁶

Principal findings are set out below.

2.1. Curated collections are enormously valuable

⁴ Core Collection Interviews. Final Report. (1 July 2004). Prepared by Alex Wright (alex@agwright.com) with: Rosalie Lack, Ellen Meltzer, Steve Toub, Roy Tennant. Available from http://www.cdlib.org/inside/projects/metasearch/corecollection_assessmentfindings.pdf

⁵ See National Science Digital Library. Focus Groups and Market Assessment. Final Report (1 July 2004). Prepared by Alex Wright (alex@agwright.com) with: Wendy Parfrey, Heather Christensen, Rosalie Lack, Felicia Poe. Available from http://www.cdlib.org/inside/projects/metasearch/nsdl/nsdl_assessmentfindings.pdf The National Science Digital Library is available at <http://www.nsdl.org/>.

⁶ Assessment for the NSDL involved a series of 5 moderated focus group sessions involving 35 librarians and teachers, drawn from educational institutions across California. Focus groups lasted between 60 and 90 minutes each and were conducted at the CDL offices in Oakland, San Diego State University, and San Jose State. Participants were drawn from academic libraries, public libraries, community college libraries, and K-12 teachers and media specialists. Assessment focusing on the core undergraduate science collection took place on the Berkeley and involved 60-90 minute interviews with a small number of students who were compensated for their participation in the survey.

Respondents from each of the user communities we surveyed were more partial to coherently organized or curated online collections than we anticipated, though for very different reasons.

K-12 teachers and media specialists are under enormous time constraints. As such, they are “less interested in ‘doing research’ than in quickly locating a few useful, high-quality teaching materials”.⁷ They are particularly interested in finding “pre-screened, selected content appropriate for a particular subject and grade level”.

The interests of *community college teachers* interests are virtually identical to those expressed by the K-12 teachers and media specialists. They want to go to authoritative and reliable collections where materials are selected to support teaching or learning in particular disciplines and at particular levels.

Academic and public librarians are enthusiastic about curated collections for slightly different reasons. The availability of high-quality, well curated, and openly accessible online collections promises to extend the holdings they offer to patrons. This collection-building impulse is particularly apparent in two slightly different ways. First, the collection-building impulse was apparent in librarians’ willingness to add virtual collections from the American West project to the pathfinders, collection guides, and other finding aids they construct to guide their users to quality information sources. Interestingly, the same collection-building impulse was apparent in the librarians’ reticence to integrate the National Science Digital Library in a similar manner. That reticence appears to be grounded in the fact that the librarians do not yet understand the NSDL’s collections aims or audiences, so they do not consider it to be reliable, authoritative, or curated. Perceived quality and brand are clearly important to librarians who continue to make informed collection decisions when pointing to openly accessible Web-based resources.

Academic and public librarians also see well-curated online collections as an opportunity to showcase the digitally reformatted materials they produce from their own analog holdings. Enthusiasm for contributing to curated collections is particularly apparent amongst the public librarians because their digitally reformatted collections are developed on a smaller scale and so take on considerably greater value when seen in the context of other related information sources. In any event, academic and public libraries may be as interested in becoming net contributors to curated collections as they are to becoming net users of them.

Undergraduate students are particularly interested in curated collections that are perceived as being reliable, selective, and authoritative. Such collections are efficient for undergraduates who are completing assignments in highly constrained time periods. Indeed, our data suggest that undergraduates so value well curated collections that they gravitate towards them and away from search engines which though efficient for searching, are terribly inefficient when it comes to finding resources that are relevant, authoritative, and reliable.

Graduate students are interested in curated digital library collections because they promise to open out broadly onto readily accessible and authoritative primary

⁷ “Documenting the American West. User Interview”, p.7.

sources which the graduate students consume like locusts in the course of their research.

Faculty's interests are little different than those of the graduate students we interviewed, with one or two important exceptions. Faculty are far more conservative in their information-seeking behavior than graduate students. It appears that many have already found and are happy with their most trusted online information sources. They use these sources repeatedly and effectively, but in a manner that impedes their interest in trying to discover, explore, and learn how to use new and highly trusted Web-based information. Any reticence is only compounded by the fact that most of the faculty we spoke with rely informally, on their graduate students to find, vet, and then point them to new Web-based information sources.

2.2. Well defined collection boundaries matter (a lot)

While the user communities we surveyed agreed about the value of well-curated online collections, they differed very considerably with regard to the kinds of curated collections they require.

Community college faculty and K-12 teachers and media specialists are particularly interested in collections that are closely tailored to their immediate curricular needs. Where primary resources are concerned (as they are with the American West), a collection's subject orientation is deemed important, but so too are collections that are organized in a way that enable teachers to quickly and easily find materials that are appropriate for use at certain educational levels (ninth-grade instruction, first-year community college). Community college faculty and K-12 teachers are even more interested in "classroom ready" learning materials (such as lesson plans and interactive instructional modules) that they can easily adapt for their own classroom use.

Public and academic librarians are more catholic in their tastes reflecting the breadth of interests inherent in the patron communities they serve. For them, a collection's reliability, authority, and integrity, is more important than its subject orientation, or content breadth and depth. Put another way, academic and public librarians seem to be looking for collections that will be useful to patrons (at least a definable subset of them) and that are narrowly defined with regard to scope, aims, and audiences.

Undergraduate students naturally enough combine the librarians' interest in authority, reliability, and integrity, with a very keen interest in subject relevance. The undergraduate assigned to write a mid-term paper on a particular metallurgical process or on the life and times of Thomas Jefferson wants to go *the* online source where they are most likely to find everything they need in support of their inquiry.

Graduate students, given their interest in unearthing seldom-used primary materials and their insatiable appetites, are perhaps less concerned with how tightly defined a collection is with respect to its scope. They want to get as quickly as they can to materials that are otherwise only accessible inside geographically disparate archives and special collections. And they aren't afraid to turn collections inside out and upside down using whatever search and discovery tools (the more the merrier) that are available to them.

2.3. Annotation is [very] important

Our assessment also demonstrated that the communities we surveyed share an interest in tools that enable very different forms of annotation.

As already indicated *community college faculty and K-12 teachers* and media specialists are particularly interested in material that is classroom ready. Accordingly, they place a premium on pathways through collections of primary sources; that is, on narrative essays that help students explore, interact with, and learn from the materials on display. They are also interested in lesson plans and in teaching materials (instructional modules) that can be slotted directly into a particular class. Although there is considerable anxiety about who ultimately is responsible for preparing such value-added content and how its preparation will be supported, there is little doubt about its desirability and value. There is also little doubt about the tool suites that are required to support such development work. Of particular significance are tools that enable users to select, annotate, and organize, materials within a collection, and to save selected materials for future use, possibly exporting them into other software environments (e.g. PowerPoint, locally maintained Web pages, etc.).

Academic and public libraries may be interested in federating some virtual curated collections with local holdings, provided they can be convinced that there is an advantage for them to do so. In such cases the libraries are likely to use whatever federation tools they have locally, suggesting that virtual collections need to expose information about their contents (whether in their entirety or as pre-selected sub-sets) in a manner that conforms to well-known metadata and data standards (e.g. via an XML gateway as an IMS content package or via an OAI server as Dublin Core records). Sub-setting – where a library treats the virtual collection selectively, federating only selected items with its local holdings – is a far more significant challenge. It relies on tools not yet available that support selectivity without relying on the painstaking, labor intensive and entirely impractical effort of “hand-picking” items that satisfy certain collection development criteria from amongst millions of digital images, texts, sound recordings, and Websites.

Graduate students and faculty require very specific suites of tools to capture and manage citations (to develop footnotes and bibliographies), and to create and manage their own personal reflections (research notes) on the materials they review. Graduate students are also very keenly interested in the most sophisticated search and retrieval tools available as only these will enable them to quickly come to terms with a collection’s contents and to discover whether it hides any gems.

2.4. Ingest may be interesting

The assessment identified a further and in some ways unanticipated challenge of librarians who were interested in adding their local holdings to the virtual curated collection. This suggested a further and potentially very important development path, notably with regard to the development of standard ingest routines – that is, routines that enable third-party contributions to a virtual curated collections. Such routines will inevitably involve technology, but perhaps more significantly, processes that enable the curator of an online virtual collection to assess potential contributions. Inevitably, such assessment will need to take place along several axes: to determine whether the offered materials fit into a collection of a given scope; whether they are of sufficient technical quality (with appropriate metadata, image resolution, assurance of persistence at a given location and through time); whether they can rightfully and legally be contributed to an

openly accessible collection; and whether they carry appropriate terms and conditions of use.

3. Developing use cases

Abstracting from the needs assessment, the project derived a number of use cases that describe in very practical terms how the content and tools resulting from the American West project may be used. The use cases also contribute to the collection planning and technology design components of the project. They act, in effect, as a bridge between the user needs assessment and the implementation effort.

3.1. Use cases for the American West collection as a reference collection

The American West project will evolve as a large virtual collection comprising openly accessible digital information bearing on the history and ecology of the American West. The collection will be extensive both in the genre of the material it incorporates (literary texts, oral histories, diaries, maps, manuscripts from archives and special collections, music and music scores, plays, poetry, databases, images, cinematography, etc) and the formats in which those materials are presented (encoded texts, digital images, encoded audio and video clips, geographic information systems). Users will have access to the collection in at least two ways. A simple Google-like search interface will support keyword searches across the entire corpus. In addition, the collection will be presented to users as a series of different browsable views (or canned searches), each of them focusing on a pre-selected set of the material organized by chronological period, geographic region, or by topic. Upon selecting a browsable view, users will be confronted with a brief narrative that explains the significance of the topic or theme in question and the types of materials that are available, and pointing to related information resources that reside elsewhere on the Web, in museums or libraries, or in canonical texts, for example.

An academic college library makes the American West available to its patrons by linking to it from several different points on its Website. The collection is available, for example, under American history, culture and society in a subject-organized list of online resources that includes licensed and openly accessible materials. Links are also available from a variety of subject guides that library bibliographers maintain to guide patrons to free Web-based resources selected for their quality and their fit with local research interests and curriculum needs.

A third-year graduate student preparing a PhD thesis on “spirituality and the American explorer” uses the simple and advanced keyword search interfaces to plunder the collection for any jewels it may contain. She finds the full-text diaries of eminent as well as lesser-known explorers. Though she has already encountered (and in most cases read) the tomes created by Lewis and Clark and other luminaries, she is delighted to discover a full text version of Patrick Breen’s diary which records the ill-fated trip of the Donner party as well as several other significant works.

An undergraduate with two weeks to prepare an essay on the impacts of urbanization in the late 19th century finds the collection through a subject-guide in the sub-categories of nineteenth century and industrialization. He goes directly to the browsable view entitled “The Western city” and enhances his understanding of the ecology of the Western city by viewing digitally reformatted stereotypes, daguerreotypes, prints, and photographs of Western cities from 1865 to 1920.

A member of the social sciences faculty in a Western community college is preparing a course on race and ethnicity in the American West and is looking for Web-based

materials that students can use to supplement the limited collections available from the college and local public libraries. She finds the American West collection in the first page of a result set returned by Google in response to a search for “American West online”. She spends a minute or two with the collection’s search interface testing the collection’s depth, provenance, and integrity, by searching for items bearing on her favorite topic: the effect on late-19th-century unionization of a multi-ethnic and multi-racial labor force in the American West. Surprised and delighted by the wealth of materials she discovers – most of them contributed by leading research libraries – she instantly recognizes the collection as a rich and trusted resource. Having sold herself on the collection’s value and integrity, she now turns her attention to the browsable views, several of which (“black America and the American West” and “Native American experience”) are directly relevant to the course she is planning. After no more than a half hour browsing through the materials in each of these views, she prepares a handful of questions for her students that will form the basis of their writing assignments. Each writing assignment refers students to reading materials that are available in the coursepack that she is having prepared at Kinkos, and to the primary materials available in these browsable views of the American West collection.

3.2. Use cases for the American West collection as an interactive and creative experience

The American West collection will be available as a ready reference resource as described above and will be used extensively as such. It will also come equipped with a suite of tools that enhance and enrich the user experience. The tool suite will enable the following functions:

- **Selection.** This feature is similar to the “shopping-baskets” that are so common in e-commerce. With it, users will be able to build and save their own collections by selecting items individually (the diary of Patrick Breen, a digitally reformatted Ansel Adams photograph), or in sets (the browsable collection entitled “Native American experience”, the entire search result set for the query “Texas missions”), saving their collections so they can be recalled and re-used at a later time
- **Annotation.** This feature will enable users to annotate items they place within their personal collections. Annotations could be supplied for individual items or for a group of items.
- **Presentation.** This feature will enable users to customize their personal collections so they have a distinctive look, feel, and functionality.
- **Export.** This feature will enable users to export personal collections so they may be saved locally and used in other preferred software.

At the Western community college, assignments based on the American West collection prove to be very popular amongst students. Some students, however, have suggested that their work would be facilitated if the browsable views to which they are assigned were introduced in a way that is related directly to the course. Using the annotation tools, the instructor writes brief essays introducing the browsable views of the American West collection that she has asked her students to review when conducting their assignments. For the “black American and the American West” view, she prepares a brief essay encouraging students to challenge the myths and stereotypes that assume the history, development, and culture of the American West stems largely from the activities of white men. In

places, the essay interacts directly with the collection, making particular reference to online artwork and literature that promulgates and reinforces this myth. The essay also touches selectively on primary evidence that is more subversive and seems to assert the role that black Americans played in shaping western culture. The essay concludes with a series of leading questions and hints at how they may be addressed by exploring particular materials in the browsable view of the American West collection. These materials are selected by the instructor as a further browsable view. Upon completing her work, the instructor saves it as a new, personalized view of the American West collection that both she and her students can access later on.

A media specialist working at a public high school is developing lesson plans in support of the American History curriculum. The high school in which she works is groundbreaking in this regard, in part because it has a skilled media specialist who is able to complete such tasks in consultation with the teachers who are responsible for delivering the curriculum and using the lesson plans. Its technology infrastructure is, however, substantially underdeveloped. Lesson plans reside on a Web server licensed from a local ISP, and comprise little more than flat or static HTML pages. The media specialist is delighted to discover the American West collection because of its contents, but also because it includes tools that enable the development of lesson plans that selectively refer to items drawn direct from the American West collection or other Internet sites. The media specialist develops a handful of lesson plans making extensive use of the annotation feature. Once the content and organization of the lesson plans is completed, she then spends some time with the “skin and slice” tool kit to ensure that the plans carry the high school’s look, feel, and brand. Once the plans are completed, they are exported as static HTML pages and uploaded directly onto the ISP Web server.

A graduate student is working on his dissertation, which focuses on the impact of geographic location in the rise of the hard disk drive industry in Silicon Valley and in various locations in east Asia. The student has been collecting citations for the last several months and uses personal bibliographic management software (EndNote) to store citations, as well as to annotate and manage them. A colleague pointed this graduate student to the American West collection, which provides access to several journal articles on the impact of geographic location in the rise of Silicon Valley. Citations with durable URLs for easy access to the full text when he wants to view them in the future, are marked in the American West portal and then exported to EndNote. Within EndNote, the student identifies and removes duplicate citations, types in a few notes to jog his memory about certain citations and continues his research. In the future, when he needs to visit the full text of the articles or when he needs to generate a bibliography in the proper format, he is able to easily perform these tasks from EndNote.

A university library has developed a renowned collection of Asian-American materials to reflect and support faculty research and teaching strengths in this area. The library links to the American West collection through a variety of its subject pages just like it does with many other online collections of free and licensed Web-based resources. Many of these collections include materials relevant to Asian-American studies. To find them, however, the user needs to search in each of the different databases (such as the American West collection, JSTOR journals). Given the library’s historic strength in this area, it launches an ambitious program to build a portal through which users can query Asian American materials as if they made up a virtual uniform collection. The portal will be built with the

library's local metasearch product (it uses Ex Libris's MetaLib software) and will reference online full-text journals selected from the journal databases to which the library subscribes, relevant bibliographic references selected from the library's online public access catalog, and relevant content from the American West collection. Using the advanced search interface, the bibliographer is able to find a large subset of the American West collection that bears on Asian-American experiences. The subset is saved as a personal collection and is made available using the export functions via an XML gateway that interacts directly with the library's MetaLib software.

3.3. Use cases for tools created for the American West project

The American West project was initiated to develop the tools that libraries and other information organizations need to build and manage digital libraries with distributed online information sources. Tools are being developed to fill a range of discrete functions as described in greater detail in Section 5. Some of the tools are already available and in use by the CDL and other libraries. Others are under development and will be released in stages over the next 18 months. The CDL will use all of these tools to build the American West collection and develop the features that support the use cases described above. It is anticipated that libraries and other organizations will also use the tools in a variety of ways to assist in their development of a wide range of digital library and online educational services. Some use cases are presented below.

Through its administration of the federally funded LSTA program, a State library has invested substantially over the past seven years in more than 200 digitization projects hosted by academic and public libraries state wide. Together, these libraries have produced hundreds of thousands of digital images, encoded texts, audio files, and maps based on vast array of unique analog materials that are stored in the special collections and archives. Although the collection comprises information dealing with a wide range of subjects, individuals, and historical events, the lion's share reflects upon the history and society of the people in the state. The problem, of course, is that the collections reside on independent Websites and are not integrated in a virtual uniform collection with browsable views of the numerous themes and topics. To address this problem, the state library provides a grant to a leading civic public library with a track record in digital library development. Under the terms of the grant, the public library is tasked to integrate these disparate collections into a virtual uniform collection bearing on the history and people of the state. The civic public library has its own core infrastructure but lacks the metadata harvesting capacity represented in the CDL tool suites. Accordingly, it acquires the CDL's metadata harvester and its curatorial tools. Since the public library serves a wide range of audiences, it also procures the annotation tool which it uses with abandon to develop numerous "exhibitions" that thematically integrate subsets of the distributed collection contents and presents the subsets with introductory essays.⁸

A federation of community colleges interested in capturing learning materials that support instruction in the basic sciences uses the collection building and curatorial tool suites to develop a virtual uniform collection based on distributed online materials

⁸ The use case is not at all unrealistic. It reflects a grant made recently by the California State Library to explore the development of a virtual uniform collection of digitally reformatted materials produced by libraries state wide with CSL support in the form of LSTA funds.

available at institutions who make course materials available online. The group has its own metadata harvesting tool suite but does not yet have the means of harvesting as selectively as it needs to in order to capture the learning materials that meet its members' specific curricular needs. Recognizing that its new learning object repository will rapidly outpace the DBMS and access service capacity available in its current FoxPro system, it is also looking to refresh its basic content management infrastructure. The group implements the CDL's XTF content management and search tool suite (available under an open source license from SourceForge) and acquires the metadata analysis, normalization, enrichment, and subsetting suite direct from CDL. With the detailed documentation from the CDL, the group is able to implement these tool suites within its local operating environment and is building its learning object repository within a couple of weeks, having expended the efforts of a single programmer for that period.

4. Collection planning⁹

The American West project will build a virtual collections by selecting relevant materials from extant digital collections contributed by eight project partners: the CDL, the Library of Congress, the Colorado Digitization Program and the university libraries at Harvard University, Indiana University, University of Michigan, University of Virginia, and the University of Washington. Nearly 70 collections have been made available to the project by these partners, comprising more than 32.5 million pages of encoded text and more than 4 million digital images, along with a small amount of sound and video files.

Collection planning has entailed three related tasks. The first task was a thorough survey of the materials on offer to identify strengths and weaknesses and to suggest appropriate ways to organize and present the materials once integrated. The second task was the development of collection building strategies that will fill essential gaps in the materials that are currently on offer, and the third task was a review of design and technical issues surfaced by the collections review.

4.1. Collection review

A thorough review of the collection content on offer identified significant strengths and candidate views through which the collection content may be presented to users.

A key strength is found in the collections' chronological coverage and suggests that it might be presented with chronological views covering the frontier period (1800-1890)¹⁰, and two sequential post-frontier periods (1890-1940 and 1940-present). Material bearing on the frontier period covers major historical events such as Lewis and Clark's explorations and travel generally (from UVA); the gold rush (from LoC and CDL), pioneer life (CDP, UW), and the railroads (CDP, CDL). Extractive industries such as mining and logging are well represented (CDL, CDP, UW), and the bulk of the collections' literary resources and material on Native Americans belong to this period.

⁹ The evaluation of the collections' strengths and weaknesses is taken in almost verbatim from an unpublished report commissioned by the CDL from Geneva Gano, UCLA. The full text of the report is available from the CDL upon request.

¹⁰ The period extends to 1890, the year the U.S. Census (and historian Frederick Jackson Turner) declared the frontier's closure.

During the first post-frontier period of 1890-1940 the population of much of the West – especially urbanized populations – doubled (or more) each decade. Many images of cities and the urban landscape are represented in the collections (CDL, CDP, UW). The non-urban landscape is also represented reflecting the birth of the modern Environmental movement, and the images it produced of wilderness even as the West was becoming densely urbanized (CDL, CDP, UW). There are also materials pertaining to the labor movement and the Great Depression (UW, CDL). The second post-frontier period, 1940-present, was marked by continued rapid population growth, a large degree of public (especially military) control of land and resources, and the growth of media and technology, including the movie industry and ultimately cybertechnology. The collections gesture toward a contemporary history of the West (UW, 's images from the World Trade Organization protests) but is most useful in the immediate post-war years where it includes at least three different views of Japanese American Relocation Administration camps, including prints by Ansel Adams and diaries and journals of detainees (LOC, CDL, CDP). Automobiles became the common American form of transportation during these years, and automobile tourism became an important industry in the West. These themes are especially well documented in the Cushman collection of photographs from Indiana and the Roads for Modern Tourists collection from CDP.

Regional, state-by-state views are also suggested, since this is a common mode of understanding place in the American West. The majority of the resources in the collections pertain to states in which contributing institutions are located, especially California, Colorado, and Washington. Materials from these three states also include the widest array of resources for social historians and educators, as contributing institutions tend to have strong collections of “daily life” images and texts. Even the Library of Congress often presents its collections by state, as in the “California As I Saw It” collection. Because the virtual collection is wide-ranging geographically and unevenly distributed (with particular concentrations in three states), a second, broader, regional designation may be useful to researchers.

Selected thematic views can also be supported and may be desirable across the period 1800-1940. This broad sweep coincides with the stereotypes, stories, and myths that many commonly associate with the American West. The topics with the most representation in this virtual collection therefore are fairly recognizable and important ones, including agriculture, exploration and travel, landscape, leisure, natural resources, pioneer, race and ethnicity, tourism, and transportation. The collection is also strong in some unexpected areas which, if they are made available to users will help expand their understanding of the West beyond traditional parameters and stereotypes. Such topics include architecture, business, politics and activism, science and technology, urban life, and work and labor. The collections are also particularly rich for social historians interested in daily life, particularly in California, Washington, and some of the states involved in the Colorado Digitization Program (viz Colorado, Kansas, Nebraska, and Wyoming).¹¹

The collections also have notable gaps, the identification of which provides a tactically efficient route map for the collection’s further development. Gaps were identified thematically but also with regard to the genre of materials on offer. Strategies for filling these gaps are addressed in Section 4.2.

Thematic gaps include:

¹¹ See the online digital library “Western Trails: An Online Journey” at <http://www.cdpheritage.org/westerntrails/>

- *Gender.* A concentration of materials by and about men and their accomplishments, with relatively little representation of women's texts or images, tends to reinforce a mythos of the West as a masculine space. Such scholars as Virginia Scharff have definitively debunked this myth, showing that the West was often a liberating space for women. More representation of women's frontier experience, women's work (possibly available through Harvard), and creative writing and art by women (such as the Cha collection through the CDL) will help to round out this area.
- *Race.* A myth of Anglo-American superiority and inevitable conquest is implicit in the doctrine of Manifest Destiny that came to be associated with the West represented most clearly in this collection (that of the trans-Mississippi West). This collection's materials dealing with race in the West are uneven, and may tend to support a narrative of white supremacy and accomplishment. Most obviously, the relative lack of images and texts pertaining to the experiences of non-whites in the West creates this impression. Major omissions include indications of African-American and Mexican-American experiences in the West, of either creative or daily life. Asian-American experience is limited, in that it is most significantly represented in tourist images of San Francisco's Chinatown at the turn of the century and in the images and texts of the Japanese-American WWII confinement camps. The Native American experience is documented quite extensively in the 19th century sections, showing a diversity of tribal groups and peoples, but contemporary representations of Native American life and creative works is almost nonexistent, creating the impression that Native Americans are no longer important in the West. Work with institutions that are specific to particular race and ethnic minority groups (but not university affiliated), may help to fulfill this project's commitment to represent the diversity of peoples in the American West.
- *Selected geographic regions.* There are two concerns with region: frontier movement and regional omissions. First, frontier movement is not well indicated in this collection. A frontier perspective of the West (that "the West" exists wherever it is indicated by a shifting frontier line) The regional omissions are most obvious. Relatively few materials have been acquired pertaining to the U.S. Southwest, including Texas. Also underrepresented are materials pertaining to the Northwest, aside from Washington state. Alaska is well represented, though in a very limited fashion, in discussions of the gold rush. Hawaii and the transnational West (including Canada and Mexico) are not well represented. These are not absolutely critical to a collection of materials on the American West, though it is important to keep in mind that this virtual collection defines "the West" by what it excludes just as much as by what it includes.
- *Sports.* Leisure is well represented in this collection, but little organized sports materials are currently available, including youth and amateur sports, local teams, and university and professional teams. By deepening the collection's materials in these areas, contemporary and non-white representation might be increased. Media clips might also help to expand genre as well.
- *Movies and cinema.* Though this is a major industry in Southern California, very few materials pertaining to this industry are included in this virtual collection. Coverage of this theme is likely to be especially problematic given copyright restrictions. At least some important progress may be made with largely ephemeral materials, for example, in collaboration with the Internet Archive.

- *Cybertechnology*. This important industry is associated with the Pacific West, but has no representation in this virtual collection. What that representation would comprise is itself a problematic question given the embryonic state of historical inquiry into this aspect of the American West. Here too, a number of paths present themselves including the development of a collection plan in consultation with faculty and other specialists who are actively conducting research in this area and collaboration with projects that are collecting materials from the dot-com boom.¹²

Genre gaps are as follows:

- *Literature*. This virtual collection is weak in its representation of major Western literature. Although the 19th century literary contributions to this project are significant ones, very little of the material contributed is actually part of a Western literature canon, nor does this virtual collection draw upon many non-canonical (popular) literary works, such as occasional poetry and dime novels. Further, no 20th century literature which makes up the bulk of Western literature, is included in this collection at all owing to copyright restrictions. Perhaps the best mechanism for filling this important gap is through a concerted digitization program that is guided by a credible canonical list of American Western literature. The program could not include in-copyright materials but would at least provide comprehensive coverage of the literature into the early 20th century.¹³
- *Music*. Three kinds of materials would strengthen the collection including scores, sheet music, and audio files (or audio clips). Models for developing such a collection can be found at UCLA and Indiana University (who are working together on collections of sheet music and musical scores) and at Seattle's Experience Music Project, which is currently featuring information about several musicians and scholars associated with the West, including Quincy Jones, X, Jimi Hendrix, and Alan Lomax. They suggest that the American West will have to extend its collection net beyond the partner libraries and into cultural heritage and academic organizations that deal in a more comprehensive way with American music.
- *Natural history*. Perhaps unsurprisingly, the library collections on offer are weak with regard to materials bearing on the ecology and environment of the American West. The Western landscape, Western agriculture, and the mining and extraction industries figure well in the materials that are available, but at present, there are limited offerings bearing on fauna, flora, geology, and the evolving ecology of the West. Here too, evidence from progress in the natural history museums community (e.g. at Missouri Botanical, New York Natural History Museum, and our own Jepson Herbaria, UC Berkeley) suggest that the American West will need to extend its collecting tentacles beyond its base of library partners.
- *Art and architecture*. Perhaps due to our sourcing collections primarily in research libraries, art and architecture are not well represented. Art libraries exist in abundance at research universities but are typically related to an academic

¹² Cf Business Plan Archive (<http://www.businessplanarchive.org/>) and DotCom Archive at <http://www.dotcomarchive.org/>

¹³ Canonical lists may be created, for example, by creating a concordance of works that are referenced in one or more credible scholarly texts. Also see footnote 16.

department rather than to a library, and thus operate as separate entities. More substantial collections may be available from the museums community but again, this will require the American West project to extend its collection efforts.

4.2. A route map for gap filling

At the third all-partners' project meeting held on October 24, 2004, the American West project broke critical ground in determining a collection building path capable of sweeping up the most significant gaps within its purview. Reflecting at length on the collection review, and the focus that it gave our collecting effort, the group agreed to commission related investigations into five key areas, notably: American literature, American music, natural history, American maps, American art and architecture. Each of these investigations is described briefly below. American literature and American music will be taken forward by the Digital Library Federation's Aquifer initiative which has resolved to focus its collection building efforts in a manner that will enhance, enrich, and extend the collection development efforts of the American South (a partnership led by Emory University) and the American West.¹⁴ The others will be carried forward by the American West project. The aim in moving forward in these five areas is that they promise to add new genre of materials (e.g. literature, music, etc.) while thematic gaps in the American West collection are filled with collection content contributed by the broader range of Aquifer and ultimately DLF research libraries.¹⁵

American literature. It is possible to identify a canonical literature of the American West.¹⁶ Yet because the boundaries around the corpus are likely to be so blurry in

¹⁴ Aquifer's collection development path is described in the report on the Aquifer meeting, held on October 25, 2004 as follows: "The American South and American West projects are making substantial progress building curated collections that bear on history, culture, and society in these two regions of the United States. Cumulatively, they are working with over 80 distributed online collections, offered by a dozen institutions and comprising millions of digital items. In shaping the collections and determining how best to present them (e.g. topically, thematically, etc) they have worked extensively with a variety of end-user communities. Through this process, the projects have developed detailed collection planning and development strategies and processes and early instantiations of the tools required to implement them. They have also identified significant gaps in the online content that has been offered to them. In consideration, then, of:

- the progress these initiatives have made already;
- the extent to which their subject focus (on American history, culture, and society) plays to the strength of Aquifer and DLF members' online digital collections (as evident through review of the DLF collection registry);
- the importance to our progress of closely defining one or more specific collecting areas as an essential starting point in all aspects of our work (technical, user-based, collection based), and
- our interest in leveraging this early work to develop the strategies, tools, and organizational and other mechanisms that will support and cost-effectively enable the prolific development by libraries and other information organizations of other, differently focused, high quality and well curated collections

Aquifer agreed to adopt as its collection orientation the enhancement, enrichment, and extension of efforts underway at the American South and American West projects. "

¹⁵ The DLF membership extends to over 30 research libraries. A somewhat dated registry of their public access online collections is available from <http://www.hti.umich.edu/cgi/b/bib/bib-idx?c=dlfcoll>. A simple search reveals literally hundreds of collections (representing millions of items) focusing in particular areas relevant to American history, culture, and society.

¹⁶ Such a list might be compiled for example in consultation with scholars and with reference to established works of criticism including Bergon, Frank and Zeese Papanikolas, Eds. *Looking Far West: The Search for the American West in History, Myth and Literature*; Handley, William R. *Marriage, Violence and the Nation in the American Literary West*; Johnson, Michael K. *Black Masculinity and the Frontier Myth in*

so many areas, it may be more appropriate to tackle American literature as a single collection undifferentiated in the first instance by regional considerations. Further, if developed incrementally and in a manner that begins with out-of-copyright materials (1776-1921) the collection development task appears tractable. Approximately 18,000 titles are in the corpus of American literature 1776-1921. At present, some 3,500 or more are available online from American West project partners Indiana University and the University of Virginia. What is required is a carefully costed plan for creating the additional titles – one that looks at the practical availability of source materials (whether printed monographs or microfilm) as well as at incremental approaches to its production. One envisages, for example, a corpus that is produced initially as a large number of digital page images that are incrementally made available as marked-up, full-text files.

American music. Here too, there is the sense that music of the American West is hard to define. Is a piece of music considered to be “Western” because it was performed in the West, because it was written and/or performed by an artist born or living in the West, because it is about the West (even if never performed there), or because the scores and sheet music were published in a Western city? Perhaps even more than American literature, the topic (and genre) begs for treatment on a national scale. At the least, a collection of American music surrounded with the full suite of curatorial tools that the project intends, will be open for subsequent subsetting and the creation of any number of regional and other views. In addition, music is complicated by the fact that it is represented in a variety of forms, such as scores, sheet music, and audio recordings. The American West project, has therefore defined a collection development path that begins with a detailed survey, conducted by an expert in the history of American music looking broadly across publicly accessible online collections of scores, sheet music, and audio files. The expert will test the hypothesis that there is a sufficiently large corpus of publicly accessible material to justify the effort of assembling it into a coherent virtual uniform collection. We are reasonably confident that the sheet music collections at UCLA, Indiana University, John Hopkins University, and elsewhere will provide a significant starting point for this important genre. A brief review of audio recordings freely available from public broadcast station Websites and of legitimate Websites that are established to support peer-to-peer audio file-sharing, suggest that there may be a substantial corpus of audio materials reflecting ethno-cultural aspects (such folk music) in particular. Here, then, the expert consultant will provide an overview of available holdings, a sense of their depth and breadth, and offer recommendations about efficient collection development strategies.

While the DLF’s Aquifer initiative develops strategies for building collections in American literature and American music, the American West project will focus its own attention (and partners’ resources) on gaps that make sense to fill on a regional basis. Working with map bibliographers, for example, it will explore the many excellent online digital library collections to determine how best to assemble a collection of digitally reformatted maps bearing on the American West. It will undertake a similar survey of the growing number of library,

American Literature; Kowaleski, Michael, Ed. *Reading the West: New Essays on the Literature of the American West*; Lape, Noreen Grover. *West of the Border: The Multicultural Literature of the Western American Frontiers*; Lyon, Thomas J. *The Literary West: An Anthology of Western American Literature*; Poulsen, Richard C. *The Landscape of the Mind: Cultural Transformations of the American West*; Rosowski, Susan J. *Birthing a Nation: Gender, Creativity and the West in American Literature*; Tuska, Jon and Vicki Piekarski, Eds. *The Frontier Experience: A Reader’s Guide to the Life and Literature of the American West*

museum, and other sites that make digital surrogates of art and architectural work available openly via the Web. While the American West project can comfortably develop strategies for building collections of maps, art, and architecture of the American West, it will engage colleagues in the natural history museums community to see whether such a strategy might be developed there, perhaps in partnership.

4.3. Design issues

The review of the collection content on offer from our partner libraries surfaced a number of issues that bear directly on the design and development of the American West collection as well as the tool suites that will support its construction and use.

4.3.1. The challenges of subsetting

A prominent theme emerging from the collection review as well as the user needs assessment is the significance of tools that enable the ready construction of different subsets of the materials in the American West collection; that is, the construction of browsable views that bear on specific periods, topics, or themes. Such subsetting tools will be essential for the CDL. End users will also use them to create and save their own collection subsets that meet their local needs and interests.

Yet the construction of these tools is substantially complicated by the enormous variation in practice adopted by partner libraries in creating item level metadata – descriptive information associated with the items in their digital collections. To assess the extent of the problem, the CDL conducted a thorough technical assessment of the collections on offer, to determine their format characteristics, their accessibility, and the metadata formats and delivery vehicles. The technical assessment revealed a number of anticipated challenges.

First partner libraries organize their materials to meet immediate local needs. This isn't a problem so long as local practice converges with those of the American West project. They rarely do. Collections at the University of Virginia, Indiana University, and the University of Michigan are particularly challenging. There, materials relevant to the American West are present in much more comprehensive collections and not immediately identifiable as a logical subset. Thus, the University of Virginia's and Indiana University's collections of online literary texts are organized by language and genre but there is no way at present to select Western literature as a subset for the American West collection. The University of Michigan, meantime, has built a huge collection of digitally reformatted materials that range from ancient papyri to 19th-century serial publications. The collection is available via a Google-like search interface and not sub-divided into categories that would enable us to tease out only those materials relevant to our needs. Even where collections are wholly relevant to the American West (as they are, for example, at the CDL, the University of Washington, and the Library of Congress), they are not readily integrated into most of the browsable views that the CDL or other users might wish to create.

As an example, the Library of Congress has offered The John H. Grabill Collection photographs, frontier life in Colorado, South Dakota and Wyoming, 1888 – 1891 (<http://lcweb2.loc.gov/pp/grabillhtml/grabillabt.html>). It is a marvelous collection comprising 188 taken by this gifted, early Western photographer. Digitally reformatted photographs are available in both high and low resolution formats and associated with richly descriptive metadata. Still, there is no straightforward means of automatically

selecting a handful of images that deal with manufacture, for example, to be included in a browsable view of the American West collection that bears on work and industry in the post-frontier period.

Second, data providers describe their collections differently. They don't all choose to provide descriptive information about the same attributes of a digital object (genre, format, and subject). Where they do describe the same attributes (such as date) they describe them differently and thus, inconsistently.¹⁷ Data providers also apply metadata at different levels of granularity. For example, in the American West collection, some digitally reformatted serial publications include metadata at the article level and others include metadata at the page level. The combined effect of these inconsistencies and idiosyncrasies is a substantial reduction in the functionality of a service that is built from items in very different digital collections. In fact, the service can function only as well as the least effective item in the collection will allow.

Subsetting is a priority for the American West collection, which means that we must have tools that will enable us to select individual information objects meet certain user-specified criteria without having manually to review all of the objects in the virtual collection. Specifications for these curatorial tools are reported briefly in Section 5.2.

4.3.2. Collection development requires a variety of strategic approaches

The review of the collection content on offer demonstrated a variety of thematic strengths. It also identified a number of gaps and forced consideration of strategies for filling them. As a consequence, the American West project has extended its collection development strategies beyond metadata harvesting to include Web crawling and acquisition of third party data and metadata for management in a CDL repository. The decision impacts directly on our development path as it becomes necessary to develop a richer array of data capture tools than had initially been planned for. The additional development work (on mechanisms for acquiring third party data content and for web crawling) will leverage ongoing investments being made by the CDL and new work that will be supported jointly by the CDL and the Library of Congress. It will enrich the American West project but will not impede its progress or add to the costs being sustained by The William and Flora Hewlett Foundation. The rationale for our adopting three collection development strategies is outlined briefly below.

Automated transfer of third-party data and metadata into a locally managed repository.

This approach results in a high degree of control over the digital information in question and a greater ability to manipulate the information (format, metadata, etc.) so as to support end-user services. The method also incurs costs (e.g. in negotiating ingest with data suppliers, processing data at ingest, and managing data it over the long term with a view to maintaining its currency) so is likely to be used sparingly. Notably it will be used to integrate:

- collections that are created by organizations (libraries, archives, museums, academic departments) that are otherwise unable deliver them online for themselves;
- collections that are created by organizations that are able to deliver them online but unable to deliver them in a way that makes them amenable to

¹⁷ For an analysis see, See Roy Tennant, "Bitter Harvest: Problems and Suggested Solutions for OAI-PMH Data and Service Providers" (2004) from http://www.cdlib.org/inside/projects/harvesting/bitter_harvest.html

- integration into virtual collections (i.e. via Web crawling or metadata harvesting); and
- collections that are at risk of loss and that consequently stand to benefit from the CDL's commitment to preserving the content that it manages locally.

This approach leverages historic strength within the CDL that has enabled us to aggregate encoded archival finding aids, encoded texts, and digital images that have been produced by a wide range of libraries, museums, and archives across the state of California into a large, uniform, online collection.¹⁸ Looking forward this approach promises to enable and empower information organizations (libraries and museums, but also academic departments) to contribute a wealth of highly specialized materials to large digital collections that take on regional and national significance. It also promises a highly efficient means by which the CDL can fill gaps in the collection with materials sourced from numerous sites. With the ability to capture and manage data content, the CDL can effectively outsource piecemeal digitization work that will need to be undertaken to fill gaps in the collection.

Automated capture of item-level metadata that is associated with data content stored, managed, and made accessible remotely by a third-party. So-called metadata harvesting is the collection strategy emphasized in the original grant proposal. It will be utilized to capture the lion's share of the content contributed by our partner libraries that will initially comprise the American West collection. Metadata harvesting will also continue as an important and highly efficient means of collection development. Having said that, the weaknesses inherent in metadata harvesting need to be understood. Using metadata harvesting, the CDL will be able to gather distributed digital information from digital libraries and other organizations that routinely create digital collections and make them available wholesale for integration into virtual or federated collections by supporting the appropriate metadata harvesting protocol. Although the domain is occupied by important data suppliers (notably, the leading research libraries) it misses out on a large number of other potentially important data suppliers that lack the technology investment in or knowledge of digital library protocols, such as smaller libraries and archives, academic departments, instructional computing centers at universities, colleges, and K-12 curriculum development offices. Finally, metadata harvesting provides the virtual collection with only a very modest amount of control over the metadata and data content that are contributed in that way. As such, it imposes very real constraints on the virtual collection's functionality as described above in Section 4.3.1. Suffice it to say here, a collection created through metadata harvesting is only as rich as the poorest data and metadata element that it contains.

Automated capture of static Web pages. Using Web crawling tools, the CDL will bring selected pages from the surface Web into its American West collection, generating basic metadata at ingest as appropriate for the discovery, location, and presentation of the stored Web pages. Although crawled Web pages are not likely to contribute significantly in the first instance to the American West collection — at present, we aim to include Websites associated with the California election campaigns, including the California gubernatorial Recall election — their

¹⁸ The materials are available in the Online Archive of California see <http://www.oac.cdlib.org/> and in Museums in the Online Archive of California see <http://www.bampfa.berkeley.edu/moac/>

contribution is likely to increase over time.¹⁹ Web crawling is highly problematic, though. It is difficult with the current generation of crawling tools to be as highly selective as a curated collection would require. Further, Web pages typically return metadata that is even more rudimentary than that which can be harvested from the least sophisticated digital library collection. Still, if conducted selectively and carefully, Web crawling affords an opportunity to capture the many wonderful expressions of Western culture and society, both historical and contemporary, that are produced and maintained by millions of faceless Internet publishers.

5. Building functional specifications for essential tools

To meet the user scenarios described above and to implement its collection building strategies, the CDL will develop a range of tools that may be categorized as follows:

- A basic digital asset management infrastructure that will provide the means of persistently managing and surfacing into a variety of access systems the data and metadata content that make up the American West collection.
- Collection building tools that will gather digital information from distributed online collections so they may be integrated into virtual uniform collections.
- Curatorial tools will select items that are gathered for inclusion in a particular collection (such as only those digitally reformatted images items having to do with Native Americans and Native American life or literature of the American West). They will include tools that enrich item-level metadata where it is insufficient to support essential selection decisions and/or the service features that are planned for the virtual collection.
- Access tools that will integrate, subset, and present collection content (however gathered), and provide basic and advanced search and retrieval functionality as well as a number of standard browsable views created and maintained by the CDL.
- Customization tools that will provide a high degree of interactivity for organizational (library) and individual users of the American West collection. With them, users will be able to select, annotate, and export items into locally defined collections that can be saved for later reference or exported into other software environments and presented with a high degree of control over the resulting look, feel, and functionality.

5.1. Basic infrastructure

The infrastructure leverages the effort underway at CDL since 2002 as part of a comprehensive re-engineering of its digital asset management capability. That effort aims to build the generalizable, scaleable, and extensible means of persistently managing and surfacing into a variety of access systems the data and metadata content that the CDL manages for the American West as well as for other digital collections. The infrastructure is highly modular, is largely based on open source components, and substantially leverages ongoing investments being made by the CDL. Essential components that are

¹⁹ For information about the Recall Crawl see <http://www.cdlib.org/inside/projects/preservation/recall/>

currently available either as production services or (with regard to the preservation repository) in beta mode are described below.

The Archival Resource Key is at the infrastructure's core. The CDL utilizes it to provide the persistent identification of digital objects. The ARK is based on a semantically-devoid identifier associated with explicit policies detailing the persistence and access commitments for any given digital object. Within CDL systems, the ARK serves as a useful index key for the location, retrieval, and manipulation of objects.²⁰

Digital Object Standard. Digital objects that are managed by the CDL must themselves comply with the CDL Digital Object Standard, currently under revision and expected to be released in a new version during early 2005. The revised standard specifies the use of the Metadata Encoding and Transmission Standard (METS) as the wrapper for descriptive and technical metadata. This version will contain specifications for images, TEI and PDF texts, and EAD finding aids. Succeeding guidelines for HTML and Web content, audio/video files, and other object types will be released later in 2005.²¹

Extensible Text Framework. The newest generation content management system at the CDL is the eXtensible Text Framework (XTF). XTF organizes and searches collections of large documents in multiple formats, providing sophisticated query capabilities and flexible navigation with search hits marked within context. XTF is based on the Lucene open source indexing and search engine, with extensive use of XSLT for configuration and displays, driven through Java servlets. It currently underpins the majority of the text infrastructure for the CDL. Collections of crawled web pages and of OAI harvested metadata are being evaluated for their ability to migrate into this new system. Although that evaluation is not yet complete, it is not unlikely that XTF will provide the core content management and search functionality for the American West collection, at least those components that rely upon digital assets that are captured and managed locally at the CDL, and those to which CDL can refer with OAI-harvested metadata. XTF has been released as an open source product on SourceForge.²²

The development of XTF is being actively pursued. Current areas for discussion include support for non-Latin languages; integration with external word databases such as defined vocabularies, ontologies, and thesauri; finer user control of search treatments such as stemming, plurals, case, and character accents; index-based spelling correction; and the expansion of covered document types. XTF, based on simple, open source components, with the ability to permit a high degree of interaction with other systems, is a model for the type of systems that the CDL envisions across a wide range of services, including important administrative functions such as content management, security, and rights management.

Digital preservation repository. In order to secure the longevity of the digital assets that it manages (including digital assets and Web pages that are captured and managed

²⁰ See <http://www.cdlib.org/inside/diglib/ark/>. In addition, ARKs and the tool suites that CDL has developed to generate and manage them are being evaluated by a variety of organization in advance of possible adoption. Those organizations include three UC campuses (Berkeley, San Diego, and San Francisco), as well as the National Library of Medicine, the World Intellectual Property Organization, Rutgers University Libraries, Internet Archive, Digital Curation Centre, New York University Libraries, and the University of North Texas Libraries.

²¹ The revised specification along with its predecessor are available from <http://www.cdlib.org/inside/diglib/>

²² See <http://sourceforge.net/projects/xtf/>.

locally by the CDL for inclusion in the American West collection), the CDL has built a digital preservation repository. The repository system is open and extensible. Automated tasks will be broken into well-defined functional components presenting clearly defined interfaces. The architecture follows the *Reference Model for an Open Archival Information System (OAIS)* published by the Consultative Committee for Space Data Systems (CCSDS). Individual components will be implemented through:

- the use of existing in-house components and wrappers;
- in-house development, leveraging existing in-house models or prototypes to the extent possible; and
- the use of third party components, with in-house developed ‘wrappers’ to integrate them into the preservation repository’s architecture.²³

Components will be loosely coupled to enable the distribution of the system across servers and to facilitate load balancing. To the extent possible, the system will be developed using languages and third-party components that are platform-independent. A database management system (DBMS) will be used for the storage and management of metadata. DBMS security will be masked through a data management service to allow for the use of different DBMS products.²⁴

5.2. Collection-building tools

Collection building tools will implement the three collection building strategies described in Section 4.3.2. and bring into a locally managed environment the data and/or metadata that will contribute to the virtual American West collection. The tool suites support the following tasks.

Automated ingest of data and metadata content. The voro suite of ingest tools permits content-submitting institutions to batch deliver objects conforming to specified standards for objects of certain types, such as EADs, texts, and images for processing and ingest into CDL content delivery systems. Voro, which is presently a set of Perl scripts, XSLT transformations, and schema validators, permits submitters to test prepared content against CDL standards, take corrective actions, and acquire a minimal set of records relating to the submittal of materials in CDL systems. The *voro* suite is being migrated into a new content management framework based on a Web services model; *voro* tools are expected to be increasingly automated and to be invoked by external systems without human intervention.²⁵ With regard to the American West project, the tools will play a critical role in capturing digital assets that are intended for local management at the CDL and possibly other data streams as well. Using them, the CDL has already built a substantial collection of digital materials (encoded archival finding aids, encoded texts, and digital images) contributed by a wide range of libraries, museums, and archives across the state of California, representing these materials in its Online Archives of California Websites. Ingest tools are continually being updated and extended so they address other kinds of information content that are currently beyond our reach (e.g. audio and video files, GIS, data, harvested metadata, and Websites).

²³ For OAIS see <http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html>

²⁴ The repository’s function, architecture, and development timeline are described in detail at <http://www.cdlib.org/inside/projects/preservation/dpr/>

²⁵ Documentation for the voro suite is forthcoming.

Automated capture of item-level metadata associated with data content stored, managed, and made accessible remotely a third-party. The CDL has implemented the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) and will be using it to federate the c.70 collections that are being contributed to the American West project by the CDL's project partners.²⁶ Briefly, the OAI-PMH specifies a method for digital repositories (also called data providers) to expose metadata about items in their collection for harvesting by aggregators (also called service providers). Using the protocol, data providers expose metadata for all of the items in a collection or for sets of items. Service providers harvest metadata from data providers, and build search and other user services on top of the resulting metadata.²⁷

Automated capture of static Web pages. Using Web-crawling tools, the CDL will bring selected pages from the surface Web into its CMS, generating basic metadata at ingest as appropriate for the discovery, location, and presentation of the stored Web pages. The CDL has established functional specifications for the Web-crawling tools and will be developing them over the next 18 months with support of the Library of Congress under the auspices of its National Digital Information Infrastructure Preservation Program.²⁸ In the interim, it uses a variety of Web-crawling tools including those provided and operated by the Stanford University Computer Science Department which is the CDL's strategic partner on a variety of projects.²⁹

5.3. Curatorial tools

These tools help to automate the process by which items gathered through data ingest, metadata harvesting, and Web crawling are selected for inclusion in a particular collection or collection subset. The tools are designed to overcome the fact that the information supplied for items in distributed and very different digital collections vary considerably in their depth and richness as described above in Section 4.3.1.³⁰

Two separate suites of curatorial tools are being developed. The first comprises support for data creators who are encouraged to adopt practices that make their collections more amenable to federation. The second comprises a suite of tools that the CDL and other service providers can use to overcome the numerous deficiencies that will remain in the metadata associated with any virtual collection they wish to build.

5.3.1. Guidelines for data creators

²⁶ For more on the OAI-PMH see <http://www.openarchives.org/>. A prototype of the CDL harvesting service – one built to evaluate development paths and so without any bells and whistles – is available from <http://dali.cdlib.org:8080/>.

²⁷ For a good example of a service provider, see <http://www.oaister.org/> which integrates item-level metadata from some 270 online digital library collections.

²⁸ The functional specification was developed with the generous support of The Andrew W. Mellon Foundation and is available from “Web-Based Government Information: Evaluating Solutions for Capture, Curation, and Preservation” (November 2003) available from http://www.cdlib.org/programs/Web-based_archiving_mellon_Final.pdf

²⁹ The CDL has evaluated a variety of web-crawling tools including the open-source suite being developed by a consortium of national libraries and the Internet Archive. Preliminary results are in two working papers available from the CDL upon request: David Kellog, "Evaluation of Open Source Spidering Technology" (June 14, 2004); and David Kellog, “Building an Archive-Quality Crawler (August 5, 2004).

³⁰ Tennant, “Bitter Harvest”

These detailed guidelines are for data creators who wish to make their content available for ingest by the CDL or other aggregators. The guidelines (<http://www.cdlib.org/inside/projects/oac/bpgdo/index.html>) have evolved over the years from the CDL's experience as aggregator of digital information created by UC and California state libraries, archives, and museums. They come replete with information for data creators about the numerous advantages (persistence, interoperability, longevity) that accrue to them through their adoption of the guidelines. To date they have been adopted by a variety of educational and cultural heritage organizations across the state, including the California State Library and the numerous digitization projects that it funds through its administration of the Library Services and Technology Act. They have also received attention nationally where they have influenced the production of best practice guides that are promulgated by the Digital Library Federation, the Research Libraries Group, and the federal Institute for Museums and Library Services.³¹ The net result is that the CDL is able to cost effectively ingest into its local content management system the hundreds of thousands of digital objects that are created by third parties and that bear largely on themes of relevance to the American West. The approach ensures that the CDL manages a very substantial digital collection that it can contribute to the American West, but perhaps more importantly, that the pipeline for the collection's further development remains well stocked with contributions from numerous archives, museums, libraries, and academic departments where invaluable materials reside that would otherwise be obscured from our purview.

As important are guidelines for data creators who wish to make their collection content available for federation via the Open Archives Initiative Protocol for Metadata Harvesting. Work on such guidelines represents the fruitful convergence of several streams of activity among a number of leading research libraries (including the CDL the University of Michigan, Illinois University, Emory University, and others) who have gained significant experience with metadata harvesting in the past few years. Guidelines, being developed in conjunction with the Digital Library Federation with support of the Institute for Museums and Library Services, will encourage data providers to:

- expose the richest metadata they have; not just the simplified (Dublin Core) versions that are minimally required by the OAI-PMH;
- expose as many metadata formats as can be managed;
- adopt consistent practices in metadata creation; and
- “wash” metadata to clean it of residual stain that results from accumulated anachronistic, idiosyncratic, or inconsistent practices.³²

These guidelines have been adopted by the American West project partners who have additionally agreed to a few more prescriptive practices to ensure that their harvested metadata content can be readily integrated into what are effectively historical collections organized roughly along chronological and regional lines.³³

5.3.2. Curatorial tools for service providers

Even with the best intentions and with the widespread adoption of best practices, the metadata contributed to virtual online collections will continue in many ways to be

³¹ See for example “A framework of guidance for building good digital collections” from <http://www.ims.gov/pubs/forumframework.htm>

³² Ibid. For more information on the DLF initiative see http://www.ims.gov/results.asp?keyword=&inst=&city=&state=55&year=9&program=gt_1009.%20gt_1007.%20gt_1006.%20gt_1011.%20gt_1012&description=on&sort=year. Details available upon request.

³³ The more detailed specification will be available from the report of the American West all partners' meeting held on October 24, 2004.

deficient. Materials that are crawled impersonally from hundreds and thousands of Websites, for example, will be immune largely to any effort, however well supported, to adopt the information community's standards. Even where data creators can be convinced about the important role that standards play, it will be impossible for any one of them to prepare material in their collections to support the full and virtually infinite array of uses to which it might be put. Finally, one needs to be careful to ensure that the barriers to interoperability are not so high that data creators don't participate purely as a matter of cost. For these reasons, the American West project has defined a suite of curatorial tools that aid in building virtual collections from disparate materials contributed via a variety of means. The tools serve four fundamental functions that are described below along with pointers to prototypes that are being developed to assess possible technical development paths.

*Analysis.*³⁴ Analysis tools will assess metadata associated with collections gathered via Web crawling, harvesting, or ingest in order to determine:

- the descriptive elements (author, title, date) that are present;
- the proportion of records that supply values for any element (i.e. populate the date field);
- the consistency with which metadata values are supplied (e.g. variant representations of names, dates, etc);⁷ and
- detect patterns of inconsistency that might guide normalization

Prototype analytical tools are available from <http://dali.cdlib.org:8080/analyze.cgi> and <http://dali.cdlib.org:8080/extractfield.cgi>. The first conducts the above analytics for the metadata elements present in a test harvest of 96,000 metadata records taken from 10 of the 70 collection on offer to CDL and returns field values as they are supplied. The second returns only the unique field values enabling detection of patterns of metadata usage (and misuse).

Normalization. These tools will standardize the way in which values are recorded for metadata elements. Tools of this type are envisaged as a suite of Web services against which metadata elements can be evaluated and through which they can be transformed to some normal representation. A prototype example is available from <http://dali.cdlib.org:8080/date.html/>. A range of data values are presented as supplied in harvested metadata records. When any one of the date values is clicked, the value is sent off to a Web service that evaluates the date and returns it in a normalized form. Although the prototype is entirely manual, it is envisaged that the normalization process would be entirely automated. Further, it can be developed in a highly modular fashion with normalization services added for dates, names, geographical places, subject headings, and integrated standard thesauruses and controlled vocabularies as appropriate.

Enrichment. When digital information is aggregated into a virtual collection, it is removed from its original context and placed into an entirely new one. This process alone can deprive the digital information of essential metadata that was never recorded but nonetheless may have been implicitly available. Thus, when harvested into the American West collection, a picture taken in 1898 of an Aleut seal hunter aboard a whale-hunting ship in Alaska is disassociated from its host collection "American Indians of the Pacific Northwest" as contributed by the University of Washington. The metadata associated with the image records a great

³⁴ A review of a particularly promising Open Source tool is contained in a CDL working paper that is available upon request. See, David Kellog, "Spotfire Decision Site" (October 11, 2004).

deal of detail, but it does not reiterate the collection's title. Accordingly, information that is potentially very useful in categorizing the image so that it may be put in a browsable view of the American West collection is lost. Enrichment is the automated process by which metadata are augmented with contextual and other information.³⁵

A prototype metadata enrichment tool is available from <http://dali.cdlib.org:8080/recall.cgi>. The prototype is based upon campaign and other Web pages associated with the California gubernatorial recall election (2004) and captured using targeted Web crawling. Web pages are notoriously deficient in descriptive metadata and so perhaps most in need of automated metadata creation and enrichment. Searching any field for Gray Davis returns metadata for 46 items. The metadata has been created automatically based on the information contained on the Web pages to which they refer. It is possible to view and/or edit the full record as automatically generated. In addition, the tool has automatically captured in each metadata record the names mentioned on the Web page to which the metadata record refers. Clicking on the name produces a new search (for the name) across the full contents of the California Recall election crawl.

Subsetting. Strictly speaking, subsetting is a function that is only applicable to metadata content acquired via the OAI-PMH. This tool will enable the service provider to determine what metadata to harvest based on criteria established by the service provider and searches based on those criteria as conducted across any of the metadata fields. Presently, under the OAI-PMH, only the data creator is able to define sets. The tool, a prototype for which is available at <http://dali.cdlib.org:8080/select.cgi>, helps build collections or collection subsets by progressively adding to (or deleting from) them through iterative queries run across an undifferentiated collection of metadata. Using the tool available in the prototype, a user might build a sub-setted view of the American West collection dealing specifically with Colorado miners, for example, by running a series of queries and then reviewing and selecting from the results those worth pursuing in greater detail. The subset as initially created can then be queried again so that the collection is progressively developed to include items that are likely relevant to the subsetted view.

Collection profiling. Each of the functions specified above, with the exception of the analysis function, will profit from a profile for specifying how a particular collection's metadata should be processed. This will enable periodic refreshment as appropriate when revisiting previously contributed collections (whether gathered by ingest, metadata harvesting, or Web crawling) for any new materials that may have been added to them. Thus, for example, when revisiting the University of Michigan's collection to gather new material (at the current rate the University of Michigan adds 3,000 digitally reformatted monographs per year to its collection), the harvested data would be massaged according to routines specified in the profile and built upon past experience with University of Michigan data. Thus, the profile would indicate the subsetting routines needed to select only that subset of new University of Michigan records that are relevant to the American West to normalize date representations, etc.

³⁵ A review of metadata enrichment/enhancement approaches is contained in a CDL working paper that is available upon request. See Michael McKenna, "Ideas about Metadata Enhancement" (April 2004).

In addition, the CDL is evaluating MetaLib, a metasearch engine product supplied by Ex Libris, as an access tool that performs a variety of these curatorial functions, notably subsetting. Detailed discussion of the product and its possible use is provided in Section 5.4.

5.4. Access tools

The world of user interface tools for content management systems is exhibiting significant churn after years of statically defined Web pages with well-defined interaction to back end databases. A key component are tools that integrate disparate content pools through metasearch engines. Metasearch tools can theoretically provide the ability to present portals of content or the ability to create subsets of available content pools for selected user communities as determined through group-based authorization decisions. Metasearch tools can also permit access systems to facilitate search parameter identification, and more importantly can assist in the identification of preferred sets of content for retrieval or manipulation. Aggregated search result sets from metasearch-enabled access systems present an ideal target for assistive aids such as ontologies, faceted browsing, format classifications, and hierarchical displays of object containers or other logical groupings of digital objects.

The CDL has licensed, installed, and is currently evaluating the MetaLib metasearch engine supplied by Ex Libris.³⁶ Although evaluation is not yet complete, it is not unlikely that MetaLib will provide the content integration and access functions required by the American West project. That is, it may provide the platform from which the CDL integrates content that is stored in separate pools or repositories into the American West collection. Again, contingent upon the further evaluation of MetaLib vis a vis XTF (see above), MetaLib may also emerge as the platform the CDL uses to store and manage the metadata content that it harvests from partner sites.

The synergy between metasearch and components-based Web services systems such as XTF is fairly obvious. One conceivable long-term trajectory for access system design then, might abstract the metasearch or other target system user interface, enabling the application of reusable components such as those in XTF (with its advanced search definition and context setting capabilities) to the determination of preferred user-based collections of objects.

5.5. Customization tools

Customization tools are being developed in collaboration with strategic partners (UCLA and the Interactive University, Berkeley) and through engagement with client libraries (such as the UC Irvine library, and the University of the Pacific library) each of which is exploring the means of providing customization tools that meet the needs of specific user communities. Our work is preliminary in anticipation of design and development decisions that will be made in the middle of the project's second year (June 2005). At present, we envisage a suite of light-weight modular tools, each of them serving specific functions as appropriate for supporting use scenarios described above in Section 3. The t that will be developed will be included on the American West Website where they can be used or acquired for inclusion in any derivative collections that users might create and export into other software environments.

Candidate customization tools most likely to be developed include:

³⁶ For product information see <http://www.exlibrisgroup.com/metalib.htm>

- curatorial tools that enable users to build their own browsable views or subsets of the American West collection;
- interface customization tools that enable users to customize the appearance of the American West collection or any subset they derive from it; and
- a suite of annotation, analytical, and export tools that enable specialist manipulation of the collection content.

Curatorial tools are described in detail above. In effect, the same tools that the CDL uses to select material for inclusion in the American West collection and its browsable view (including tools that assist in analyzing, normalizing, and enriching metadata) will be available to end users who wish to prepare their own custom views of the collection.

Interface customization tool kit. Among the customization tools, the most progress has been made here. The tool is used to customize the look and feel of a collection subset or view as may be created by a user. Colloquially, we refer to the tool as “skin and slice”. This functionality is enabled through the use of XSLT-based configuration files that enable intermediary organizations to claim CDL-provided services as locally provisioned. The CDL’s interface customization tool-kit is available from <http://www.cdlib.org/inside/diglib/repository/customize/> along with detailed documentation pertaining to its use. These tools are currently being used by the CDL to customize coherently organized subsets of the digital objects stored in its own repository (interfaces for the Online Archive of California and Museums in the Online Archive of California). They will be centrally important to the American West project as they will enable third party customization of the the look, feel, and functionality of the American West collection or any subset thereof.

Specialist tools are being evaluated, designed, and developed as follows.

- *Annotation tools* will enable users to sort, organize, and provide their own description of items and collection subsets, and/or create research notes, lesson plans, or exhibitions around a selected set of collection content.
- *Citation management tools* will enable users to identify, parse, capture, and export citations in a format that allows direct linking from the citation to the online version of the object (journal article, bibliographic record, etc) to which it refers. This tool is being developed to serve the needs of graduate students and faculty who are particularly interested in exporting bibliographic citations into their own citations databases where they can be activated as live links to the underlying object where it exists online.
- *Export tools* enable users to capture individual collection items or groups of items (as collection subsets) and any annotations they may have supplied around those items and export them in formats appropriate for use in other local software platforms, e.g. as HTML files for use on local Web pages, as PowerPoint files, as citations appropriate for loading into EndNote and other bibliographic software, or as IMS content packages for import into IMS compliant learning environments, etc.