

REALIA Project White Paper

REALIA Project Managing Board: Jeff Overholtzer, Scott Siddall, Alan Boyd, Arlene Forman, Jeff Barnett, Sara Suelflow
December 2002

Summary

The essence of the REALIA Project is expressed in its mission statement:

The REALIA Project will develop and implement a searchable digitized media database which will provide instructors of modern languages with teaching resources accessible via the Web. REALIA Project will publish royalty-free, faculty-reviewed media for scholars and students, and be open to all disciplines that wish to contribute or use cultural materials appropriate for instruction at the college and secondary level. The REALIA Project seeks to increase through collaboration the quantity of high-quality teaching and learning materials by providing a respected venue for media projects that otherwise might not be shared or published.

The project name has two meanings. Of course, one refers to signs, posters street scenes and other representations of everyday life that have long supplied the raw material for teachers of modern languages. The other meaning comes from the acronym REALIA: Rich Electronic Archive for Language Instruction Anywhere. The faculty leaders of the consortia represented in this project have long had a vision for using the Internet and other technologies to provide a searchable database of teaching materials accessible anywhere, any time. December 2002 marks the end of a one-year, initial phase in exploring realization of this vision. The REALIA Project will announce in early 2003 a Web-enabled collection of about 200 images contributed by seven faculty members and representing two languages. The Project leaders, a combination of faculty members, technologists and librarians, have chosen technologies and cataloging standards and crafted policies to ensure the collection meets high pedagogical standards and will accommodate growing numbers of contributions from diverse contributors. The project leaders also have addressed legal and promotional issues and seek to expand the collection in a framework of scholarly discussion on model uses of media in the teaching of modern languages and cultures.

Medium-term (six months) and long-term plans are developing to guide development of the project, pending additional funding. Surveys and other means of gathering feedback will help ensure the project meets the needs of its constituencies. In addition, project leaders have held discussions with organizations engaged in complementary efforts, and is open to partnerships.

History

The idea for a web-based media archive devoted to language teaching came from faculty participants in ALIANCO (Allied Languages in a Networked Collaboration Online), the modern language collaboration of the Associated Colleges of the South, who met at the University of Richmond in April of 2001. Members of this group met again in August 2001 at the ACS Technology Center to discuss desiderata for the archive. Subsequent presentations of other media projects by participants from GLCA, (Great Lakes Colleges Association), ACM (Associated Colleges of the Midwest) and CET (Center for Educational Technology at Middlebury College) led to discussion of how a central archive might be beneficial for the institutions of the tri-consortium (ACS, GLCA, and ACM) and other potential collaborators. After developing a mission statement, the joint group created task forces to find designers for the prototype and members for a preliminary editorial board. The REALIA Project then received support and input from the Global Partners Language and Technology Task Force. This task force, consisting of members from all three Global Partners consortia (ACM, ACS and GLCA), initially met at the Center for Educational Technology (CET) at Middlebury College in June of 2000. During this meeting the task force discussed various possibilities regarding the use of technology in language instruction and how such instruction might better serve students planning a study abroad experience, as well as those remaining on campus. Particular consideration was given to this effort in regard to the Global Partners overseas centers in Kenya, Central Europe (Czech Republic and Russia), and Turkey. During this meeting the task force developed a survey for distribution to language faculty on all Global Partners campuses. The survey gathered information on current technological resources available for language teaching and asked what resources faculty would like to have. Results of this survey indicated that faculty were keenly interested in increased availability of electronic images. These similar interests and overlapping goals led to partnership and joint planning with ALIANCO for the REALIA Project.

After the August 2001 meeting, members from the three consortia embarked on the next two steps of the project: 1) creating a survey to evaluate the needs of language faculty and availability of existing image collections and 2) creating a team to lead the project for one year and oversee the creation of a working prototype.

The managing board

The project team or Managing Board, consisted of two content specialists (language faculty), one librarian/archivist, one instructional technology specialist, and two specialists with expertise in database administration and web delivery. In addition, the project would receive staff support from the ACS Director of International Programs, and an instructional technology specialist at the ACS Tech Center. Members of the Managing Board are:

- (Project Coordinator) Jeff Overholtzer, Washington and Lee University, Director of Technology Education
- Alan Boyd, Oberlin College, Associate Director of Libraries
- Sara Suelflow, Macalester College, Web Coordinator
- Jeff Barnett, Washington and Lee University, Spanish faculty and Director, Program for Education in Global Stewardship

- Arlene Forman, Oberlin College, Russian faculty
- Scott Siddall, Denison University, Assistant Provost for Instructional Resources and Director of Instructional Technology

The Managing Board met face-to-face twice during 2002: February 7-10 and September 19-20. Both meetings were at the ACS Technology Center in Georgetown, Texas, and included these additional participants: Teresa Wise, ACS, Atlanta, Director of International Programs; Suzanne Bonefas, ACS Technology Center, Director of Technology Programs; Deena Berg, ACS Technology Center, Instructional Technology Specialist; and (by phone), Maciej Ceglowski, Web Applications Developer and Systems Support Specialist at CET, who consulted on various technical issues. The Board supplemented these meetings with frequent electronic communication and phone calls.

Consultation

In its deliberations, the Board has always chosen a collaborative approach. It has sought counsel from numerous sources in order to gain useful ideas and to explore mutually beneficial partnerships. Those consulted have included:

- Louis Janus, director of Program For Less Commonly Taught Languages, Center for Advanced Research on Language Acquisition.
- Max Marmor, director of collections development for ArtStor, a Mellon-funded initiative to develop a digital image collection focusing on fine art.
- Nancy Millichap, director of MITC, as well as other consortial members in GLCA, ACM, ACS.
- Members of the Consortium of Liberal Arts Colleges, a group of 59 highly selective liberal arts colleges.
- Christina Updike and other leaders of the Madison Digital Image Database (<http://cit.jmu.edu/mdidinfo/>), a fine arts image database created at James Madison University. Viewing the project raised several questions, including what type of functionality is required for language teaching, where the image is a means for generating discussion, versus fine arts instruction, where the image is the end in itself.
- Ross Scaife, director of SUDA On-Line, a web-based translation of the Byzantine encyclopedia known as the Suda (<http://www.stoa.org/sol>). Based at the University of Kentucky, this is a collaborative translation project that allows for browsing, assignment, submission and vetting of text.

The Managing Board also reviewed many other online archives of media resources, and generated its own survey to help gather additional information. The detailed survey of modern language faculty members in Global Partners institutions produced information that was instrumental in guiding the project. In addition to affirming the importance of media in teaching – 91 percent of the 115 respondents agreed with the statement, “Images play an important part in my teaching” – they commented on:

- The ways in which they use images in their language courses
- Their current sources of images used in teaching
- Their willingness to provide images for use in a database for educational use

- The strengths and weaknesses of two early examples of media databases developed at Washington and Lee University and the Center for Educational Technology. (Details: survey questions, <http://pixie.cet.middlebury.edu/survey/index.php>; survey response data: http://www.colleges.org/~alianco/Alianco_media_prototype/survey_reports/countsLanguage.html)

Information gleaned from the survey, and consultation with outside experts and other research efforts were used to craft a plan for REALIA Project. The plan addressed the needs of the target audience – teachers of modern languages – and also gave consideration to technology, cataloging, marketing, organizational and other issues.

Accomplishments

As of December 2002, the REALIA Project has yielded these results:

- The name REALIA Project was chosen and the Internet domain “realiaproject.org” was registered for a two-year period.
- The Managing Board decided to create a prototype tool with 180 images contributed for two languages - one using the Roman alphabet, the other employing a non-Roman alphabet. Spanish and Russian were chosen to provide the REALIA Project team with a representative set of cataloging and other challenges that might be encountered when working with a large number of faculty members and diverse languages.
- Six faculty members - three in Russian and three in Spanish from the GLCA, ACM and ACS consortia – were chosen as charter contributors for the project. Late in the year, an additional faculty member in Spanish was solicited to be a contributor. The images – totaling around 200 – contributed by the faculty members will be available in the REALIA Project database early in 2003.
- Technical standards were created for the images to be used in the on-line database (see appendix 2)
- Selection criteria were established by faculty members on the Managing Board. A key component of the REALIA Project is the peer review of visual materials for the collection, and these criteria will help ensure that materials in the database are useful to teachers of modern languages (**see appendix xx**).
- Guidelines were created for faculty and students gathering media for use in the project (see appendix 1).
- Slides, prints and other visual materials submitted by faculty members were scanned in conformance with the technical standards designated by the Managing Board. Some faculty members submitted already-digitized materials.
- Software was chosen for the on-line database after a thorough consideration of commercial, open source and home-grown solutions (see “Technology” section below).
- Legal documents were created to address privacy, copyright and other issues (see appendices 3, 4 and 5).
- Cataloging guidelines were developed for images in the collection. The guidelines were developed jointly by librarians on the Managing Board and the faculty members who served as charter contributors. The guidelines follow standards such as Dublin Core in

order to allow precise searching, interoperability with other collections, the eventual incorporation of map interfaces (using technology known as geographic information systems) and advanced search technology such as Latent Semantic Indexing (see “Cataloging” section below).

- A Web site for the project was created, including a logo, descriptive pages and an interface to the database of images.

Cataloging

The REALIA Project database structure was developed with existing and emerging metadata standards in mind. Following a review of the faculty survey results and discussions at the first meeting of the Managing Board, Alan Boyd prepared a draft metadata standard (see appendix 6) for board review and faculty prototype testing. Since the scope of the material to be included in the database was very broad it was decided that three widely accepted standards would be followed:

1. The Dublin Core (<http://www.dublincore.org/documents/dces/>), with qualifiers, was chosen as the basic structure for the data fields.
2. Library of Congress Subject Headings (<http://authorities.loc.gov/>) were selected as the basic descriptive thesaurus.
3. The Getty Thesaurus of Geographic Names (<http://www.getty.edu/research/tools/vocabulary/tgn/index.html>) was selected as a source for geospatial terminology and coordinates.

The newly promulgated NISO draft standard Z39.87-2002, Data Dictionary--Technical Metadata for Digital Still Images, was also consulted at a later stage of prototype development.

In the summer of 2002 customized data input spreadsheets and instructions were field tested with the initial group of six faculty prototype contributors. Fundamental lessons were learned quickly during the early phases of prototype metadata assignment:

- Time constraints were a major hurdle for faculty contributors.
- Contributors were able to use the recommended Getty site for geographical descriptors
- The assignment of even a single, basic LC subject heading (i.e., the REALIA type field) was too confusing for contributors to accomplish without more training.
- Descriptions and pedagogical use statements provided by the faculty contributors did provide rich and valuable contextual data for each image.

At its September 2002 meeting the Managing Board approved a simplified metadata structure (see appendix 7) and decided to advise faculty to focus their attention on title, descriptive, and pedagogical use fields. It is hoped that this will make much more efficient use of faculty time. More expansive faculty-composed free-text in these latter two metadata fields will also allow the REALIA Project to experiment with new semantic retrieval algorithms under development at the Center for Educational Technology at Middlebury College. (see further discussion on Latent Semantic Indexing in the Technology section below).

The Managing Board also began exploring methods by which other tasks needed to fully assign metadata could be accomplished. There are both routine tasks, which could be done by student assistants, and more complex tasks, which will need to involve a core of library staff consultants. If the REALIA Project is able to develop a distributed method of engaging faculty, student assistants and library staff in high quality metadata assignment it could well serve as a model for similar projects operating on both the local and inter-institutional level.

When the Board's concurrent investigations led to selection of CONTENTdm as the software platform on which to develop the prototype database, our Dublin Core formatted metadata was quickly loaded and fully supported on the new system based on a the direct mapping of REALIA Project metadata fields to the internal CONTENTdm Dublin Core standards (see appendix 8). CONTENTdm's XML export capabilities will in turn allow REALIA Project to migrate to other standards-based software platforms in the future should that need arise. In addition, XML export, along with use of Dublin Core standards, offers the promise of interoperability with other databases – for instance, an umbrella database that would encompass the REALIA Project and other collections.

Technology

Despite the fact that the “people issues” are the most important aspect for this project, nearly every aspect of project management, cost, staffing and faculty use depend on the selection of technology in some manner. We focused our attention on two primary areas of technology: the hardware and software needed to support the REALIA Project experimental prototype and the hardware and software needed to operate a public, production server for the Project in the longer term.

The technical phase of this effort began with a general assessment of four different approaches for licensing software for the REALIA Project prototype and production systems. We considered homegrown software, commercial (“off-the-shelf”) software, open-source (freely distributed) software, and a partnership with other organizations that already were operating digital asset management systems. Cost-effectiveness was always a concern in our selection of technical approach and ultimately we elected to move forward with a commercial solution.

In general terms, homegrown software solutions can be designed and implemented to meet a very high percentage of a project’s customized needs. The tradeoffs include the need to pay for the software’s design, development, documentation, support and long-term maintenance. A comparison of homegrown and commercial software considerations is provided in Table 1.

The advantage of open source software is that it is freely distributed without licensing fees, however we did not find an open source application that met most of REALIA Project’s needs. Support for open source applications comes directly from the community of software developers and when the number of developers for a particular application is not high, support for the application can be limited or worse. We did not identify any digital asset management applications in the open source market with a widespread and large developer community behind it. James Madison University’s Digital Image Database was particularly interesting but it was at the time not being published under an open source license.

We considered how the needs of REALIA Project could be met through a partnership with an existing organization. While such a partnership could provide many of the Project needs, it was clear that we would have to comply with metadata standards, import and export practices and other workflow standards that were already established by a potential partner. For example, we considered an alliance with the Digital Media Center at OhioLINK (<http://www.ohiolink.edu/dmc/>). The Managing Board felt that we needed more flexibility than these opportunities would afford, especially during the prototyping or experimental phase of our work.

The Managing Board's experience with commercial offerings was augmented by reviews and in some cases on-site testing of several commercial programs including Alchemedia, CatTrax, CONTENTdm, Cumulus 5 Workgroup, Destiny, Extensis Portfolio, Gallery Systems EmbARK, Informix Media Management, and Luna Imaging's Insight Software Systems. Most of these systems were not designed with the specialized needs of educational uses in mind; they are geared more toward advertising agencies and publishing houses that need to control access to assets. Others were too expensive to be considered.

In an effort to get started on the experimental phase of the project quickly, we chose Extensis Portfolio as the tool to distribute to charter contributors of images. The costs of educational licenses for this personal digital asset management tool are relatively low (less than \$100 per copy), and the Project's metadata standards could be incorporated into the Portfolio schema.

At the same time, it became clear that CONTENTdm, a digital asset management system developed initially at the University of Washington, was a strong contender for the central server-based program the Project would use in the long term.

A decision was made to distribute a Microsoft Excel template with prescribed metadata fields as a tool for gathering image metadata from faculty contributors. The images associated with each row of metadata were carefully named and provided as separate files. This process combined with the import process of CONTENTdm provided us with the system for creating the REALIA Prototype. Thus the Project did not use Extensis Portfolio as an input method but rather we conflated the technologies for the prototype with that of the production system. The REALIA Prototype currently operates on a CONTENTdm server based at the Associated Colleges of the South Technology Center.

Importing images for the Prototype

The spreadsheet approach for importing a large number of initial images into CONTENTdm was successful. The faculty contributors were already familiar with this software, thus minimizing training needs. The spreadsheet along with the image files (each associated by name with the corresponding metadata) were distributed to the Managing Board members who tidied up the metadata and ran a "batch import" of the data and images through the CONTENTdm "Acquisition Station."

The Acquisition Station software is locally installed (on the desktop computer) software that is used by faculty contributors and collection administrators to add, delete and edit images and associated metadata. The software communicates with the central server at the ACS Technology Center (see Figure 1).

The only shortcoming of the spreadsheet import method involved the importation of fields of metadata containing Unicode characters to support Russian and accented characters in Spanish. The current release of CONTENTdm (3.3) does not fully support Unicode characters; they can be input into text fields (using the editor in the Acquisition Station) but the indexing process ignores them. Additionally, during a batch import process, the Unicode null character can be entered, which truncates the entire metadata record and prevents the image from appearing in the CONTENTdm database displays.

Otherwise the spreadsheet import method is acceptable, however direct input through the Acquisition Station software is always the preferred method both for large batch importation as well as routine collection administration.

Routine collection management

Faculty contributors should have the CONTENTdm Acquisition Station installed on their desktop computer, and they should be assigned two accounts on the server: one web server account for interacting with the web interface of the collection and one system account (usually the same username and password) that gives them FTP access to upload their materials.

The faculty member specifies their collection when logging into the Acquisition Station and then can add images from existing files or an attached TWAIN scanner. Images must adhere to the Project REALIA image standards; there are standards for both full resolution images (1600 X 1200) and service images (if the service image is not generated automatically by the Acquisition Station software). In any case, the thumbnail is generated automatically. Metadata is edited into the form-based Acquisition Station and the project saved on the local computer until such time as the images and metadata are deemed ready for transfer to the server. Once the images and metadata are transferred to the server, they are held for review and release by a collection administrator of the Project.

Strengths and weaknesses of CONTENTdm

CONTENTdm provides many features that are useful for the management of Project REALIA. Most notably these include

- the ability to distribute the preparation and submission of images and metadata.
- a queue which permits administrative review before images and associated metadata are added to a collection.
- the ability to handle any media format in addition to images (sound and video clips, PDF files, etc).

Useful features which serve the public who access the collection include

- the web interface for public browsing of images.
- both collection browsing and search tools, including Boolean logic.
- the ability to display web-optimized service images while making full resolution images accessible as needed.
- The ability to display multiple perspectives of a single object, for instance, the sides, top and bottom of a chair, or the front and reverse sides of a coin. ContentDM's "compound object editor" permits the collection manager to associate multiple views of an object so that they are conveniently grouped in search results

Two features which are needed by Project REALIA are full support for Unicode characters including indexing for searches, and a Macintosh client for the Acquisition Station (although the spreadsheet method provides a reasonable interim solution). CONTENTdm has promised addition of these features in early 2003.

Features for the future

The Project REALIA prototype links each service image (640 X 480 resolution) to a full resolution (1600 X 1200) image that can be used for printing, detailed examination and other uses. In addition, CONTENTdm incorporates the industry-standard "Mr. Sid" image compression routines which allow one image to meet purposes of both a rapidly delivered service image as well as a detailed full resolution image. The Project will pursue the licensing and use of Mr. Sid server technologies for future work.

A particularly vexing challenge is the amount of faculty time required to create useful metadata. Project REALIA is interested in pursuing a developing technology called "**latent semantic indexing**" as a means to address this issue. LSI was described in 1990 (see <http://lsi.research.telcordia.com/lsi/papers/JASIS90.pdf>) and the Center for Educational Technology has been actively working on implementing this technology. "The method examines a document collection as a whole, to see which other documents contain some of those same words. LSI considers documents that have many words in common to be semantically close, and ones with few words in common to be semantically distant." [http://javelina.cet.middlebury.edu/lisa/out/lisa_intro.htm]. Project REALIA contributors could enter for each image a much larger volume of natural language data related to the image, and this large volume of less organized metadata could be indexed using LSI methods. CONTENTdm can export metadata in XML format which may improve the opportunity to integrate LSI methods into the metadata search routines used in CONTENTdm. Overall, such an approach might improve search results while changing the requirements of our faculty to labor over the creation of precise metadata.

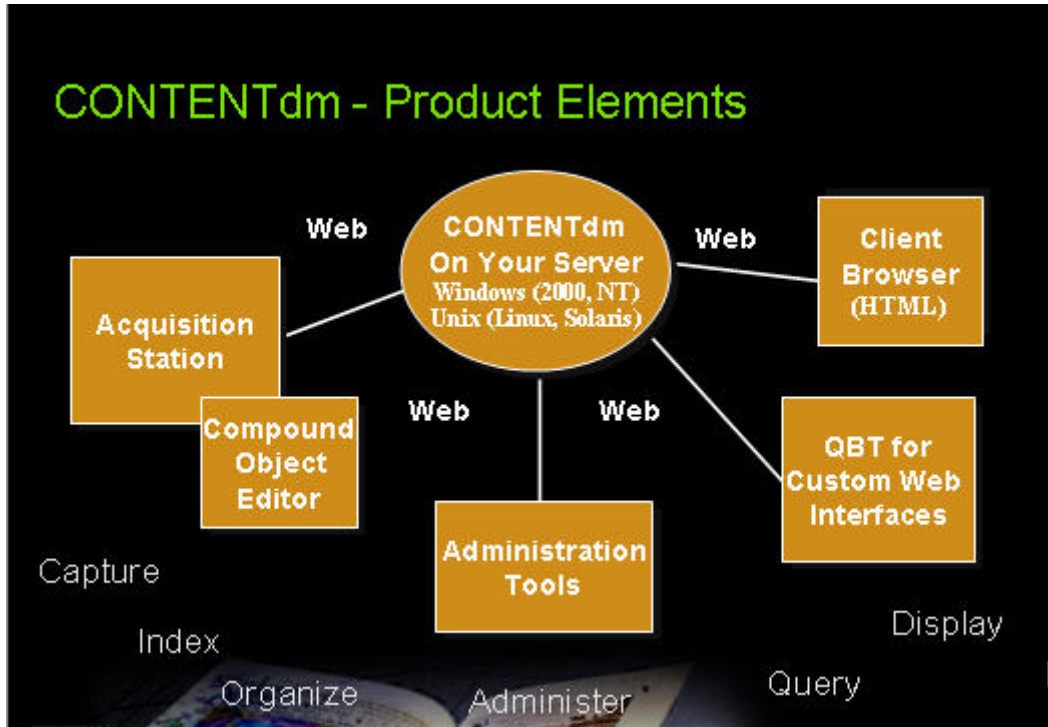
Table 1 Comparison of REALIA Project Software Approaches

DEVELOPMENT ISSUES			
Option	Tasks	Pros	Cons
<p>Customized system (homegrown, with open source components)</p> <p>Example: MySQL database running on a Linux server; programs coded in PHP.</p>	<p>We take on responsibility for designing, programming, testing the application.</p>	<p>We control the design of the interface, including changes in response to faculty feedback. Typical homegrown applications meet 90% of user needs.</p>	<p>We must design the interface, select appropriate industry standard components. We must commit human and capital resources to hiring and supervising programmers. We must ensure that the code is maintainable (documented since programmers will come and go). We must test and debug the application in a production environment.</p>
<p>Partnership that provides access to comprehensive system</p> <p>Example: OhioLINK's Digital Media Center</p>	<p>Selecting the system/partner.</p>	<p>No human or capital resources used to develop the application. A higher ed partner's system may meet 90% of the users' needs since the goals are more specifically geared to faculty.</p>	<p>We can suggest changes and new features to the interface or operation of the selected system but they may or may not be implemented. We endure the effects of bugs until the developer resolves them.</p>
<p>Licensing of commercially marketed comprehensive system</p> <p>Example: CONTENTdm</p>	<p>Selecting the system/partner.</p>	<p>No human or capital resources used to develop the application. Typical off-the-shelf products meet 80% of the users' needs.</p>	<p>We can suggest changes and new features to the interface or operation of the selected system but they may or may not be implemented.</p>

Table 1 (cont.) Comparison of REALIA Project Software Approaches

SUPPORT ISSUES			
Option	Tasks	Pros	Cons
Customized system (homegrown, with open source components) Example: MySQL database running on a Linux server; programs coded in PHP.	Maintain the software (bug fixes, new features, maintain compatibility with underlying technology that changes frequently).	We would be in control of the timing for bug fixes and new features.	We would have responsibility to maintain the software (maintain and manage programming staff and assure quality controls). This is an on-going cost.
	Provide ongoing user support through a helpdesk (phone and/or e-mail).	Control over the support process.	This is an unpredictable and probably costly responsibility.
	Create and maintain documentation for system administration and end users.	We could customize the users' documentation.	Documenting a homegrown application is crucial and time-consuming.
	Provide a training program for end users.	We're in the best position to do this.	None? We should be training users in any case.
Partnership that provides access to comprehensive system Example: OhioLINK's Digital Media Center	Report bugs & needed features to technical support contact. Prepare customized documentation based on that of partner.	Someone else maintains the software and documentation.	We endure the effects of bugs until resolved by developer. New features may not be implemented as fast as we want or at all.
	Provide static, first level technical support (web pages, how-to publications). We would act as single point of contact for technical support from partner.	Partner provides technical support to the project (inquiries aggregated by us). Partner may provide second level technical support for advanced questions.	We may have to pay an annual maintenance fee for technical support.
	Provide a training program for end users.	We're in the best position to do this.	None? We should be training users in any case.
Licensing of commercially marketed comprehensive system Example: CONTENTdm	Report bugs & needed features to technical support contact. Prepare customized documentation based on vendor version.	Someone else maintains the software and documentation.	We will pay an annual maintenance fee on the order of 15-25% of licensing fee (i.e., \$1,000/yr for CONTENTdm)
	Beyond basic training, we could provide web instructions and how-to publications but vendor would provide most technical support for each campus.	Someone else provides most support.	We pay for this in annual maintenance fee above. Exceptional support may incur additional charges.
	Provide a training program for end users.	We're in the best position to do this. Some vendors provide "train the trainer" programs.	None? We should be training users in any case.

Figure 1



Future

The first-year prototype phase of REALIA Project was designed to provide experience in technical, logistical and other issues that would guide possible long-term growth of the project. This experience has helped us identify medium-term and long-term goals necessary to fulfill the mission of the project.

Six-month goals

The current board will serve during a transition period of approximately six months, guiding the project in achieving these goals:

1. Testing the Web interface to the database using an on-line survey, as well as focus groups in three participating consortia.
2. Participating in a scholarly dialog concerning the use of media in teaching. This goal can be advanced in the short term by participating in a conference on digital image databases sponsored by the Midwest Instructional Technology Center at DePauw University.
3. Increasing the quantity of cataloged images in the REALIA Project by sponsoring a workshop for modern language faculty in summer 2003 at the ACS Technology Center. The workshop would include invited faculty members in Russian, Spanish and perhaps other languages, who would be assisted by librarians and technologists in editing and cataloging their images.
4. Promoting the REALIA Project by visiting institutions in the ACS, ACM and GLCA consortia to explain the project and recruit faculty, contributors and editors as well as technologists and librarians.

Long-term goals

Future directions and ideas anticipated for the project, include:

- Broadening the project to languages in addition to Spanish and Russian.
- Forming language-specific editorial review boards to vet media submissions to REALIA Project. Adding media such as sound, video, and QuickTime VR to the database.
- Adding search technologies such as Latent Semantic Indexing (LSI), which provides “intelligent” searches of natural language entries (see “Technology” section above)
- Providing a stream of conferences, workshops, and other opportunities for faculty to produce media and discuss pedagogical uses.
- Identify and promote exemplary use of media in teaching modern languages and cultures.
- Explore mutually beneficial partnerships with groups and organizations engaged in complementary efforts.

These efforts will be informed by continuous feedback from our audience: professors of modern languages. We will measure use of the media database and associated Web pages through analysis of server logs, and will conduct periodic surveys to obtain detailed critiques of the services provided by the REALIA Project.