Quality Management in Digital Libraries

Vamshi Ambati, Pramod Sankar, Lakshmi Pratha, C.V Jawahar

(Regional Mega Scanning Centre, International Institute of Information Technology, Hyderabad, IND-500019)

E-mail: vamshi@cmu.edu, pramod_sankar@research.iiit.net , lakshmipratha@iiit.net, jawahar@iiit.net

Digital Libraries have received wide attention in the recent years allowing access to digital information from anywhere across the world. They have become widely accepted and even preferred information sources in areas of education, science and others. When the full potential of digital libraries is realized, any citizen will for the first time be able to access all human knowledge immediately from any location. For a successful and useful Digital Library, the assurance of quality of the product from digitization process is quite essential. In this paper we discuss the major quality concerns of data from the digitization process, and how we assure for better quality before they are web enabled for the end-user to use. Also we discuss a Quality Management Framework that we applied for the improvement of both the Quality of the digital output and the efficiency of the processes.

Key words: Digital Library, Quality Check, Quality Assurance, QOS

1. INTRODUCTION

Digital Libraries have received wide attention in the recent years allowing access to digital information from anywhere across the world [2][6]. They have become widely accepted and even preferred information sources in areas of education, science and others [1]. With the increasing awareness of the benefits offered by Digital Libraries and the increasing understanding of know how to capture printed documents in digital information[3], we now are moving into a phase where we need to address other concerns and issues of digital libraries which are not so obvious. Quality management, among others is one of the most important of the issues of the day. Digital Libraries are not only a way of providing rich features to anybody from anywhere around the world, but also a means of preservation of content for the use of future generations. Hence preserving quality is of utmost importance.

In this paper, we attempt to focus on the important quality problems faced in phases of content development and web hosting in a large scale digital library. We share the experience from the developing content for the Digital Library of India project [4] and also propose the quality metrics and management process that were established over the course of the project that aim at producing an end product of high quality.

The Digital Library project (DLI) of India started book digitization in the year 2002 with a major objective of preserving the rare and age-old collection of books related to Indian History, Geography, Culture, Agriculture, medical sciences, technologies, philosophy, epics among others, by means of digitization. The project, motivated by the ULIB project [8], also aims at making these precious works of various authors, available online for students, research scholars, and any common person. Since the world is now in the *Go Global and Internet Era*, the emphasis on online reading and having digital library has become the latest trend and technology for many educational institutions. The project attempts to digitize and preserve rare pieces of the vast collections available in India and making it accessible to every one free of cost. The selection of these books follows the standards of the copy right law, pertaining to the country of the source libraries, and along with the willingness and the support from the authors for publishing it online.

The DLI project started with few scanning centers and now due to the increasing number of scanning locations these centres are grouped under different Regional Mega Scanning Centers (RMSC). The increase in scanning locations and the centres marks the increase in the digitized content. Around 14 million pages of digitized books were delivered in a span of two years and were web enabled for view to the public. There was a standard increase in the digitized content on a daily basis that reached the Regional Mega Scanning Centres for Quality Check. The initial procedure that we have applied for quality check is a simple manual check and it helped when the amount of data turned over for quality control is far below the count of today. This massive amount of data being turned over on a daily basis for quality control lead us to have a serious concern on automating the quality control. The issue of maintaining

the specified scanning and processing standards for the digital output, removing the duplicates that are scanned across the centres and within a centre, validating the outcome of the data after OCR, started to occupy our major effort towards the process of Quality Assurance. The major problem we faced was the massiveness of the data which slowly became unmanageable. This led us to apply a quality management framework for the better quality of the digital output and the web enabling process.

The rest of the paper is organized as follows. In section 2 we discuss the nature of the threats of Quality and also discuss a few such important problems. Section 3 discusses our Quality Management Framework that attempts to address most of the issues and thereby enhancing and ensuring the quality of the outcome. Section 4, 5, 6 discuss each of the important parts of the framework in detail and we conclude in section 7.

2. ISSUES OF QUALITY IN DIGITAL LIBRARY OF INDIA

2.1 NATURE OF QUALITY ISSUES

In large digitization projects, with collaborative work and distributed efforts of various parties involved in the process we often find compromise of quality due to which errors creep in. As with the case of any project, the resources involved in the DLI are humans and machines. Hence, the errors can be broadly classified into two categories which we explain below.

Manual Errors

These are errors that arise mostly due to miscommunication between human resources or due to their incompetence or non-adherence to process and standards. These are the most important and perhaps the costliest of the problems that could be seen. For the entry of metadata of a book, we largely rely on the librarians for the accuracy and credibility. But as we are well aware and as the experience shows, this is a very error prone process and also a biased process. For example, librarians might not be well advised about the hierarchy and ontology of book classification and so might classify a book as belonging to a category of "Art" when it really belongs to a category of "Music". The result would be that the book does not show up to the end-user who is interested and searches for "Art", but might show up as a false search to a person searching for "Music". An even worse situation could be assigning a completely irrelevant category.

Similarly the process of scanning and producing digital content needs to be dealt with proficiency and care. An improper scanning operation made without following the standards set, may not be useful and suitable for an end-user. These problems are unreliable and manifest in various forms. Manual errors made cost a lot to the project as firstly there is no better way of identifying the errors when data is generated in such massive scales and secondly because the data thus generated is not useful to end-users, destroying the purpose of the project.

Machine Errors

These are errors that creep in due to the understandable limitations of the software or improper configurations of the machines and software. Data generation which is done during the scanning is the key phase as we obtain digital images from the books. Once the images are obtained, we would have sent the book back to the library and so any problems that need the book to be re-obtained and re-scanned are costly. For example, an image processing algorithm that checks for the 'Skew' aspect in images may have certain limitations in detecting it and could classify a non skew image as containing skew. However such a problem could be rectified with better versions of the software whenever available. However errors due to improper configuration of scanners during the scanning phase are more serious as the data generated becomes less useful due to low quality.

Finally, as most of the books scanned are very old, they often have the problem of missing pages, damaged pages and missing title pages etc. These have to be carefully separated out in the initial stages, as it would be difficult to distinguish between such problems and problems caused due to manual or machine errors.

2.2 QUALITY ISSUES AND CHALLENGES

Incomplete and Incorrect Metadata

Most of the books scanned in the DLI project are procured from sources like libraries and government archives and hence contain metadata entered by knowledgeable personnel which can be relied upon, but is still debatable due to individual biases. However a major portion of the sources of books in the project have metadata only in non-digital formats and so these have to be fed manually. This process though inevitable is understood to be prone to errors. Due to these varied sources of book flow in the DLI project in multiple languages and due to the lack of standard

formats, metadata is missing, incorrect or incomplete or sometimes difficult to interpret. Inaccurate metadata hinders fruitful search and retrieval of books, categorization. For example, if a language field of a book is entered as English, when the book actually is French, then the possibility of the book being shown up in searches is completely lost. This problem if occurs with 'Titles' or 'Author' names would make it difficult to verify for duplicates.

Book Duplication

As mentioned earlier, due to the varied sources of books, like libraries, government organizations, institutions and personal collections those are distributed across various parts of the country, duplicates could arise between scanning locations maintained by an RMSC and also across different RMSCs. Effort put into scanning a book, processing the images and quality assurance can not be afforded to be spent on duplicates. Communicating metadata across centers and within scanning locations is important. The Duplication of the books can be identified only using metadata of a book like the title, author, publishing year, edition, etc. However, if the metadata is incorrect, missing or incomplete as discussed in the previous section, it makes the duplicate detection all the more difficult.

Image Quality

The books scanned in the DLI project are very old. Some books that have been scanned are as old as 150 years. In situations like these, the quality of the images that are produced by scanning does not appeal to the user to read. Yet another problem is improper scans. When data is scanned and book pages are generated at the rate of 3000 pages per day, the problem of an improper scan and irregular image becomes unpreventable. Though ideally the scanners are of six-sigma quality, the operators operating need to take care of the configuration and the parameters of scan.

Text and HTML Quality

Optical Character Recognition (OCR) is the module that reads an image, understands the textual content in the image and outputs the text. Text is the appropriate means of storing data, occupies lesser memory than images, is easily editable and helps in indexing and searching of the documents. The OCR is thus a very crucial aspect of the digital library.

However, the OCRs are not 100% accurate, due to the various limitations imposed by the underlying recognizer. Let us assume that a page consists of 30 lines each containing 30 words, and a word on average consists of 4 characters. A typical page would contain 3600 characters. In case the OCR is 99% accurate it results in 36 erroneously converted characters per page and over a book of 500 pages, there would be about 18000 errors. Such a situation is unacceptable to the user, and it also reduces the efficiency of the index and search. Given that the OCR limitations have been reached we look at alternate approaches to correct the OCR output.

3. QUALITY MANAGEMENT FRAMEWORK

In order to address the quality concerns mentioned in the earlier sections, we have established a Quality Management Framework at the RMSC Hyderabad and this setup has given promising results. In this section we explain the process established by Digital Library Project and then introduce the phases of a QMF that has been experimented in the project. The digitization process can be divided into the following phases:

a. Pre-Scanning Process: Process of entering metadata of a book and getting approval for scanning.

b. Scanning Process: After approval for scanning, the contractors scan the books at their respective scanning locations.

c. Post-Scanning Process: Consists of cursory image processing and conversion to textual formats. A quality control phase is also advised after scanning and this is carried on by the contractors.

d. Web Enabling Process: Digital Library receives the digital content in all formats and performs quality checks and hosts them online.
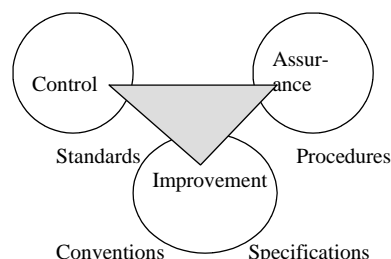

Figure 1: Quality Management Framework

To ensure high quality in each of these processes, a set of standards and procedures need to be put in place. Based on data collected from the manual quality checking, a Quality Management Framework (QMF) was built, motivated by traditional Software Quality[5], that can identify most of the errors encountered hitherto. Automated tools were developed to support the framework. The QMF for the Digital Libraries consists of the following components:

Quality Control: Quality control deals with establishing a set of standards, conventions and processes in every phase of the digitization process. These activities are preventive and suggestive measures to fight errors and promote quality. It also includes measures for quality required to be followed by the contractors who carry on the scanning operations.

Quality Assurance: Quality Assurance is performed on the end product received at the Regional Mega Scanning Centre, unlike the quality control which is performed continuously in the process and on intermediate outputs, this component of the QMF is basically a verification and validation of the product and the process applied during the quality control phase. The defects found in the digitized books during the assurance phase are sent back to the contractors for correction.

Quality Improvement: Quality Improvement is performed on all aspects of the project. Most importantly the final output from the digitization project is improved by state of art research activities supported by the participating institutions. Also the existing procedures and standards, tools and techniques are improved based on the feedback and evaluation of the deliverables. In the following sections we also discuss the efforts being undertaken for enhancing the quality of metadata, HTML and OCR in particular.

The QMF resulted in better Quality of Service (QoS) and improvement of QoS. The following sections explain each of the components in QMF in greater detail.

3.1 QUALITY CONTROL

As we have stated earlier, the challenge of managing the massive data that gets generated on a day to day basis became one of the major concern for the QMF to address. Accordingly the conventions and standards are made rigid and robust through out the project, in order to ensure completeness and correctness of the digital product. In this section we have highlighted the standards and the process that are to be applied that ensure quality control. The scanning centers and operators are instructed to follow these quality standards and processes. Tools are provided that can be tuned to certain quality parameters which are required for them to maintain through out the phases of the project.

Standards

To start with, achieving uniformity and consistency in the data delivered in the project was the top priority. Hence we set the conventions and standards for the scanning input parameters, image processing parameters and also the naming conventions of folder, file structures and file formats. This helps in the easy access for the web enablement and the content delivery, and also maintains uniformity in the product outcome among all scanning centers.

The identification of a book among our distributed and large information systems requires us to ensure the mandatory application of a definite and unique identifier with readable naming conventions. Barcode naming convention for the book is defined as
<Centreno>_<VendorNo>_<Locationno>_<Source Library>_<Bookno>
Eg: (2 03 004 010 0000001) means
2: RMSC_South (Mega Scanning Centre)
01: Par Informatics Pvt. Ltd. (Vendor)
004: IIIT Hyderabad _Main Campus (Scanning Location)
010: State Archive (Source Library)

It is also required that proper folder structure is maintained while submitting the data for web enablement. The digitized content of a given book is stored in image and text forms. The images are stored both pre-processed and post-processed in folders named OTIFF and PTIFF, the format of storage being ".tif" files. The text is stored as plain text, in rich text format, and as a html page in folders named TXT, RTF, HTML respectively. In the case when an OCR is not present for a language, the text folders will not be present in the book folder. The higher level details such as book title, author, publisher etc. are stored in a metafile in XML format stored as "meta.xml".

The processed images of the scanned books are the backbone of the entire DLI process. They are a valuable test-bed for research and development in various areas. The parameters of a processed scanned image are:

1. Dimensions: All pages in a given book should be of the same size as given by the page's Width x Height. In cases where a diagram or map occupies and entire spread, the width of such pages could be double that of other pages, but the height should still remain the same.

2. Dpi: Dots per Inch is the resolution at which the page is scanned and specifies the number of pixels in the image that represent an inch in length of the page. Every page should be scanned with a resolution of at least 600dpi.

3. Compression Algorithm: The compression algorithm should ensure lossless compression, while reducing the disk space required for storing the images. The compression algorithm to be used is the "CCITT 4 Facsimile" scheme.

4. Margin: Each page should have a uniform margin of at least 300 pixels on all sides of the textual content.

5. Skew: Skew is the term used when the text in an image appears slanted. This happens when the book is either tightly-bound or loosely-bound which makes the spread non-uniform. A maximum of 2 degrees of skew can be allowed in the scanned images.

6. Blank Pages: Blank pages in the book should be identified and annotated. This would help saving the bandwidth required for transmitting the blank pages, if the user prefers not to view blank pages.

These standards and conventions have been evolving and new standards are laid out to avoid any un-handled exceptions as and when they are encountered.

Process

The digitization of a book starts with an expert librarian entering the regular metadata for the books that need to be scanned. The metadata is first uploaded onto the DLI portal hosted at the RMSC for checking of possible duplicates from elsewhere at other scanning locations. However due to a continuous flow of books from libraries all over, a significant overlap is expected not only between scanning locations but also across RMSCs. Hence the uploaded metadata has to be synchronized with the other RMSC databases and then duplicates are detected in the uploaded metadata. This ensures prevention of duplicates in the system. An automated tool was developed at RMSC Hyderabad where contractors submit the meta-data of the book to be scanned. This data is checked against the database of books that are already scanned or are being scanned at a different location. Books are cleared for scanning only when there is no duplication of effort. The duplicate density or the number of duplicates has reduced to a large extent based on the metadata verification.

In case of Indian language books, the metadata is entered in a customized ITRANS format (IT3) developed by IISc and Carnegie Mellon University. This scheme is used to transliterate Indian languages to English. This ensures that the meta-data entered is standard and well defined titles are given for Indian language content. The verification of the metadata helped reduce the number of duplications to a large extent saving valuable resources for the project. The graph in Figure 2 shows the reduction in the number of duplicates encountered in the digital library.

The contractors then carry out the scanning operations on these books. The tools and commercial software used in the project can work with tunable standards and configurations and so the job of contractors is to ensure the right configurations are plugged in. Quality control on the contractors end, is probably the most important as the digital content is generated from the physical books here. After the scanning phase there is a rigorous quality control handled by the contractors to check errors in the early stages. We realize and support the fact that an error seen in the early stage is worth a lot more than discovering the same after enabling on the web.

3.2 QUALITY ASSURANCE

The Quality Assurance phase ensures that the configurations, conventions and standards defined by the DLI project are adhered to. Any discrepancy is reported so that appropriate action can be taken to correct the mistake. It also identifies the areas where the standards are being violated so that steps can be taken to ensure that such errors do not repeat. This phase also provides vital statistics for the over all quality improvement.

One of the major concerns addressed during this activity is the quality of images and the OCR output is high. This is because the end user has the option of viewing the digitized content either as an image or a

recognized text. Thus, the quality of images and the OCR output has to be very high. Also the OCR output should be properly displayed in the layout, as in the original document. The submitted digital books are verified for the quality of the contents to be ascertained.

Apart from the folder and file naming conventions given in Section 3, it has to be ensured that the files are properly named and each page in the book, including blank pages and covers, has a corresponding file in each of the directories. No stray folders or stray files can be allowed in the book or its sub-folders. To ensure that the book is complete and free from errors before uploading onto the web for public use, it is checked for the presence of all the required folders and files. This is done automatically by a tool developed at RMSC-Hyderabad, called QualCheck. Any book that has missing files or folders is sent back to the vendor, for it to be resubmitted with all required items.

The QualCheck tool can automatically check for each of the parameters defined for the images, and generates a report for each of the errors. The report is stored in XML format in a user specified location. A snapshot of the tool is shown in Figure 3.
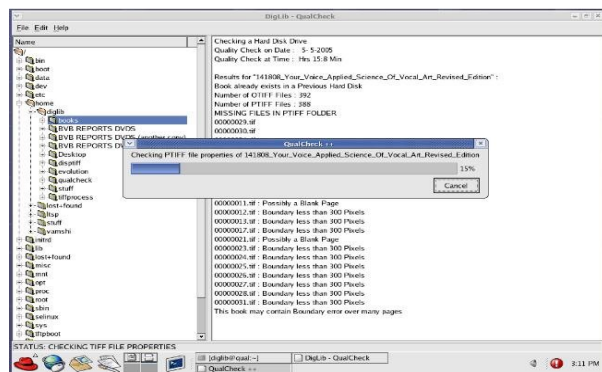
Figure 2: Automatic tool for quality verification of a digitized book.

The assurance phase has given us interesting statistics and also helped us assess the progress achieved by following the QMF. The quality metric used is the *number of defects per book*. The defects are further categorized as duplicate pages, incorrect and incomplete metadata, damaged pages, and pages that do not adhere to the defined scanning and processing standards. Figure 3 shows data accumulated over the past year and it shows that the defects gradually reduced by a large scale and we reached our minimum by the first quarter of 2005.
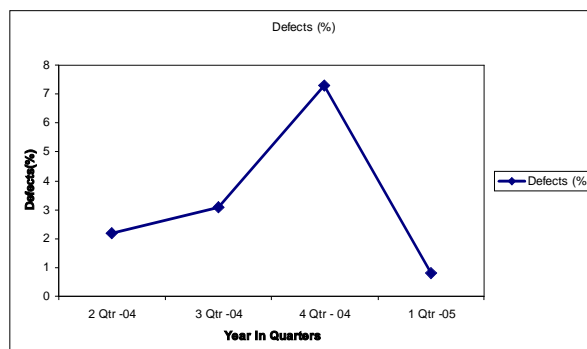

Figure 3: Reduction in number of defects

3.3. QUALITY IMPROVEMENT

Though quality of the product is ensured and enhanced by the Quality control phase, we still have scope for improvement in various aspects of the system. Reputed institutes are involved in the DLI project, which have been working towards building solutions for constant improvement by processing the final product from the digitization phase. The following are a few measures we have resorted to. There is a huge research that needs to be put in this direction, as there is plenty of scope for further improvement. Only rich and quality content, would appeal to end-users and this in turn would decide the real success of any digital library.

Improving Quality of Metadata

Most of the books scanned in the DLI project are procured from sources like libraries and government archives and hence contain metadata entered by knowledgeable personnel which can be relied upon, but is still debatable due to individual biases. However a major portion of the sources of books in the project have metadata only in non-digital formats and so these have to be fed manually. This process though inevitable is understood to be prone to errors. Due to these varied sources of book flow in the DLI project in multiple languages and due to the lack of standard formats, metadata is missing, incorrect or incomplete or sometimes difficult to interpret. Inaccurate metadata hinders fruitful search and retrieval of books, categorization and at the same time brings in scope for duplicate entries of the same book.

In order to improve the quality of metadata, we resort a mix of machine and manual techniques. We first correct frequent spell errors, and formats of the metadata and then work with librarians for manual

further correction. Since books are scanned from multiple locations and libraries through out India, we need librarians who can understand different languages and have a diverse knowledge on various subjects. Usually the metadata is verified by remotely distributed librarians who can log into the system and make necessary corrections over the web. Automated identification and correction of metadata using statistical techniques and human feedback has also been yielding successful results.

Another significant improvement is we promote the concept of structural metadata concept proposed in[7] for a book object in our digital library. This metadata contains information pertaining to each page like the size of each page, whether the page is blank, or has an important context attached to it - like the beginning of chapter, end of chapter, index, preface, table of contents etc. Such information enables us to improve the navigation of the end user through the book and also improve search and retrieval systems. RMSC-HYD provides tools for entry of structural metadata. This is guaranteed to improve the user experience in browsing and finding information.

Improving Quality of OCR

Different techniques have been proposed to correct OCR output based on the knowledge available regarding the language. Such knowledge is available in the dictionary of the language. Each word output by the OCR is checked for its presence in the dictionary. If the word is not present in the dictionary, the reverse of the word is checked against a reverse dictionary. This enables to pin-point the location where the OCR could have made a mistake in converting a particular character.

Another method could be to build a probabilistic model that encapsulates the distributions of pair wise characters. This model could provide the possible characters that could follow the recognized character. This allows us to tune the OCR accordingly and improves the accuracy of results to a great extent. However, this method would require a lot of training data.

Improving Quality of HTML

The OCR output is stored in plain text as well as in html formats. The html file can store the layout information of the content of the page. This allows the user to view the page as it was printed. To enable this feature, the commercial readers usually hard code the coordinates at the sentence level and build a table where each line is coded for its font type, size and width-and-height. However, all browsers do not have the same font configurations. The hard coding of coordinates can lead to overlap of text and image zones. Also the text would not be free flowing as the display is at sentence level. An example is shown in figure 4.
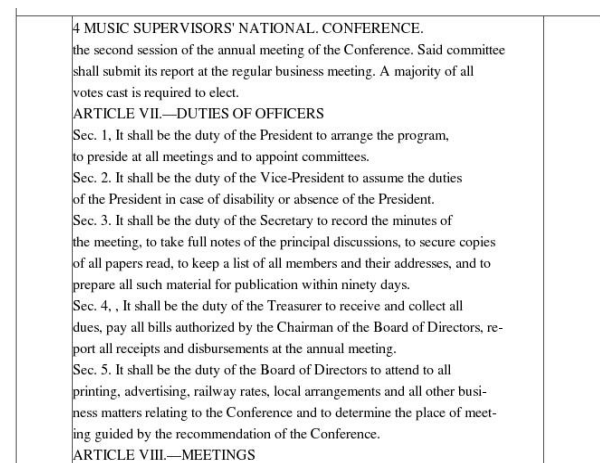

Figure 4: Misaligned and non-free-flowing text

To overcome this problem the HTML page could be reconstructed on-the-fly with change in font size and type. This can be achieved by dynamically creating tables with the hard coding being at word level. This ensures that the text is free flowing and looks justified to its space. Also this method preserves the integrity of text blocks and image blocks. An improved HTML output is shown in figure 5.

Figure 5: Free flowing text in presence of two columns and displayed in a large font size

Here the two column text is displayed in a large font. The text can be observed to be free flowing and devoid of any overlaps.

4. CONCLUSION

We have discussed the issues and compromise of quality that arises when digitization projects grow to a massive scale. We explain the scenarios from the experience of executing a large scale, highly distributed and extensively collaborative project, the Digital Library of India. We discuss the Quality Management Framework and its components that have helped us reduce the duplicates, errors and improve the quality of the end product for a better quality of service and the maximum benefit of the users. The QMF would be further refined based on the user feedback and reviews.

5. ACKNOWLEDGEMENTS

REFERENCES

Gary Marchionini, Hermann Maurer.: The roles of digital libraries in teaching and learning Communications of the ACM, Volume 38, Issue 4, April 1995

Michael Lesk, Understanding Digital Libraries, Morgan Kaufmann, 2004

David Bainbridge, John Thompson, Ian H.Witten: Assembling and enriching digital library collections Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital library ies,2003

Lakshmi Pratha, Vamshi Ambati and Pramod   Shankar: Content Development for Digital Libraries: Regional Mega Scanning Centre Hyderabad, Digital Libraries Spe cial issue on the VishwaBharati July'05

William E. Lewis, Software Testing and Continuous Quality Improvement, Auerbach, 2000

Alexa T.McCray., Marie E. Gallagher.: Principles for Digital Library development. Communications of the ACM Vol ume 44, Issue 5(May 2001)

Naomi Dushay.: Localizing experience of digital content via structural metadata Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries,2002

The Universal Library Project: http://www.ulib.org