

Ensuring Access to Mathematics Over Time: Cooperative Management of Distributed Digital Archives

**A collaborative project of
Cornell University Library and Göttingen State and University Library**

1.0 ENSURING THE LONGEVITY OF DIGITAL JOURNAL LITERATURE

Librarians have traditionally accepted responsibility for the long-term preservation of the scholarly serial literature they have purchased. This has been realized through the dedication of resources to cover the binding of journal volumes, conservation of materials over time, as well as by maintaining adequate shelf space to house them. In the past, journal publishers have not contributed financially to these preservation activities. The growth of digital serial literature has presented librarians with many complex problems in fulfilling their familiar archival and preservation functions. Libraries tend to license rather than own electronic literature from publishers. Shall we depend on publishers to maintain long-term electronic access to their archives when they have not played this role for print literature? Or shall libraries, working cooperatively with each other and with publishers, meet the challenges of developing digital archives? In addressing the myriad of questions surrounding how best to develop and maintain reliable digital archives, this study will focus on a complex discipline. We will develop an archive of serial mathematics literature that will be available to libraries worldwide and at the same time serve as a model for similar efforts in other disciplines within the library and publishing communities.

The Cornell University Library (CUL) and the Göttingen State and University Library (SUB) will create a distributed, interoperable system for the long-term preservation and dissemination of digital serial literature in mathematics and statistics. To do so, we will develop and implement a system that adheres to the principles put forth in the Open Archival Information System (OAIS) Reference Model. At an operational level, we will establish metadata requirements and communication protocols for data ingest, data management, archival storage, archive administration, and access processes. The two institutions will establish data and metadata structures that will enable separately administered databases to function as a digital archiving system, and the preservation requirements of this distributed set of repositories will be managed through common processes.

Digital mathematics content offers unique opportunities for digital preservation research due to the complexity of its document objects. Scholarly communities and publishers that depend on mathematical expressions in their literature rely heavily on the typesetting language TeX (mathematics, statistics, physics, chemistry, etc.). The TeX language is highly variable and is designed to allow author-defined macros. As well, TeX documents often have multiple components or associated files. Since TeX files encode underlying typesetting instructions that require further processing for direct viewing, a presentation version of such articles also needs to be captured. The most common presentation file formats are currently PostScript or PDF, and more recently DjVu is increasing popular, particularly in Europe. The proposed system, therefore, must manage complex archival and presentation formats from multiple sources.

The architecture and administration of the proposed system presents challenges as well as benefits due to the distributed nature of the system as conceived. The project will develop administrative protocols whereby both institutions have the capacity to administer the system, verify its contents, and repair any deficiencies. In building an implementation of the administrative layer in OAIS, the project will develop an interoperable and generic model capable of being administered technically and managerially by multiple partners. This distributed administrative capability will be reflected in the system architecture through multiple access points and a common administrative interface. Creating duplicative access to the administrative interface allows further expandability of the system, enabling other administrative partners to join over time.

The need to develop and implement functional, integrated metadata element sets for archive administration and resource discovery and access is critical at this stage in digital library research. The participants recognize this need and seek to build an operational data and metadata repository for long-term storage and retrieval of digital objects in this subject domain. In exploratory discussions of such an archival system, the participants have met with representative content owners, including professional societies (EMIS - The European Mathematical Information Service), large commercial publishers (Springer-Verlag), and small, independent publishers and distributors (Cornell's Project Euclid, see section 5.1). These discussions have been conducted in the context of the Electronic Mathematics Archives Network Initiative (EMANI), an international effort to address challenges faced in the digital archiving of mathematics. Additional participants of EMANI include Tsinghua University Library, Beijing, and Orsay Mathematical Library, Paris. Preliminary agreements have been reached with participating publishers who will provide journal content at no cost to the project proposed here.

In summary, project objectives (discussed more fully below) include the following:

- Using OAIS as a guiding model, define the requirements for an interoperable, distributed preservation and dissemination system.
- Develop the system architecture and workflow that integrate distributed repositories into a single preservation and dissemination system, including the rationales and means for exchanging content and its associated metadata.
- Define the metadata elements necessary for the preservation and dissemination system.
- Build a working prototype that implements the developed design requirements for a preservation and dissemination system. Content for the prototype will be a selection of resources from participating publishers.

Both partners have long histories in digital library research, including a combination of theoretical and practical experience. More specifically we have worked together in the area of distributed repositories, and we are leaders in the development of common practices and standards for preserving digital content. This project will allow us to focus on critical research questions:

- What are the functional requirements of a distributed interoperable digital archive? What are the functional requirements for user access and for administrative management of the system?
- What is the nature of interoperability in a distributed digital archive? What level of interoperability is required?
- What are the criteria for evaluating digital archiving and interoperability? What criteria indicate successful archiving?
- What metadata element sets are necessary to meet the functional requirements of a digital archiving system?
- What communications protocols are required to meet the functional requirements of a digital archiving system?
- What conditions need to be met at each node of the distributed archive for it to work successfully?

The final stage for developing a comprehensive digital preservation program involves establishing inter-institutional collaboration and dependency based upon a secure integrated matrix of digital archives. Ultimately, users should be able to trust that a digital archive will provide secure, reliable access to resources over time without constraints based upon geographic location, language preferences or requirements, or subject discipline. Much of the ongoing work in OAIS-based projects has involved single implementations, generally at one institution or within a homogeneous technical environment. This project will contribute to the contextualized interoperability that will enable such a matrix. Using this implementation as a proving ground, the project will demonstrate the means and mechanisms for developing and managing a distributed digital archive that instantiates the OAIS reference model.

2.0 DIGITAL ARCHIVING

2.1 OAIS Reference Model

Since its publication in 2001, the Open Archival Information System (OAIS) Reference Model has received a great deal of attention in the archiving and preservation communities, and it has become the dominant and accepted model for planning and implementing digital archives [OAIS; Lavoie; OCLC, 2002]. The aerospace domain, represented by major space agencies in North America and Europe, led the development of OAIS with active input from United States National Archives and Records Administration (NARA), the CEDARS project (Leeds, U.K.), and other library and archives programs and initiatives. The abstract nature of a reference model provided the ideal means for non-domain specific coordination, discussion, and planning.

OAIS is modular, scalable, and flexible, making it ideal for use as a model. In order to build a working implementation of OAIS, however, one needs to take the model's high level functional requirements and articulate them at a much more specific level. Various organizations and projects have begun this work, and there are several projects that have implemented portions of the OAIS model, including CEDARS [CEDARS] and NEDLIB [NEDLIB].

The OAIS Reference Model provides the basic framework for digital archives planning at Cornell University Library (CUL) and Göttingen State and University Library (SUB). Key features of the model (shown in Figure 1 below) include the Submission Information Package (SIP) which provides metadata for the ingest process, the Archival Information Package (AIP), and the Dissemination Information Package (DIP), which provides tools for accessing and disseminating resources stored electronically.

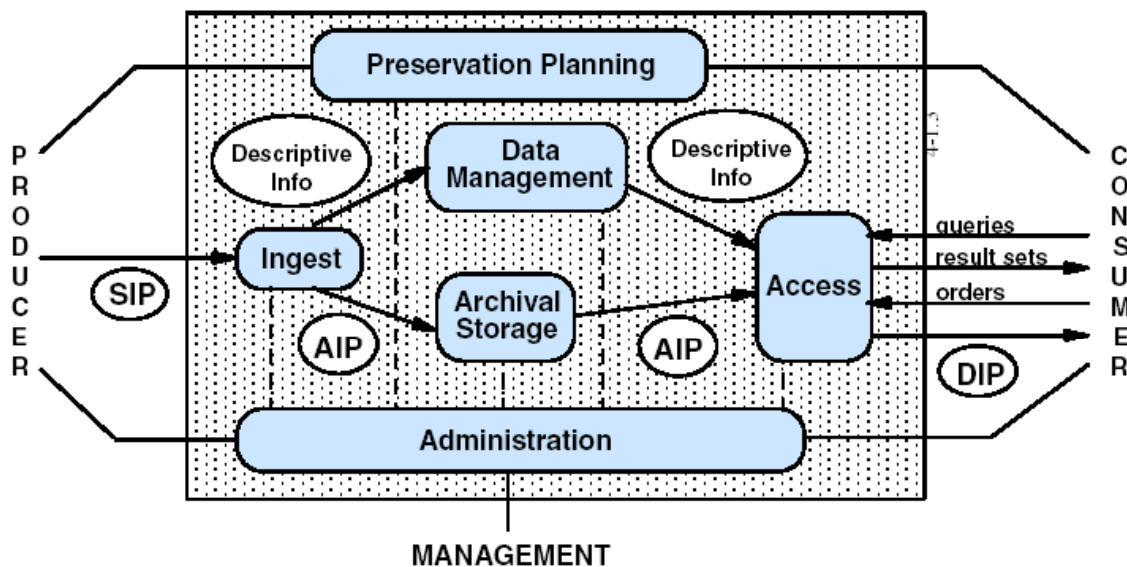


Figure 1: OAIS Reference Model

In terms of Submission Information Packages, both partners are gaining experience in the area of ingest processes, as well as increasing understanding of the metadata required for effective long-term management of digital content. As part of work on Project Euclid (<http://projecteuclid.org>) and the Physics arXiv (<http://www.arxiv.org/>), discussed in section 5.1, CUL has made considerable progress toward developing ingest procedures and the descriptive metadata requirements necessary to fulfill the functional needs of an access system. In its work on the CARMEN project (<http://www.mathematik.uni-osnabrueck.de/projects/carmen>), SUB has investigated various metadata requirements for digital resource management, including archiving.

The implementation of the Archival Information Package will involve the design and construction of a distributed system for long-term preservation of digital content. Aspects of the distributed nature of such an archive have been investigated as part of the NSF-DFG supported Mathematical Monographs project (<http://www.library.cornell.edu/mathbooks/>) between CUL and SUB. This project is concerned with distributed searching across heterogeneous repositories, where the semantics of search queries must be abstracted and communicated via a common protocol. A similar logic will be applied in the current project. Although different local repositories and systems are in place, the ability of these systems to perform preservation analysis of every partner's content is key.

In addition to this project, CUL has invested considerable resources in planning a campus-wide archival system. This Common Depository System (CDS) represents a shift from a project-based

approach to a fully implemented digital preservation program (<http://www.library.cornell.edu/iris/dpo/cds.html>) and is synergistic with this project. For over a decade, Cornell has been creating digital content that is of long-term interest to a wide user audience. Beginning in 1999, with an Institute of Museum and Library Services (IMLS) grant entitled "Preserving Cornell's Digital Image Collections: Implementing an Archival Strategy," CUL began working on a comprehensive solution to capture, preserve, and enable long-term access to digital collections at the University. The initial project focused on image files (for the project report and other documents, see <http://www.library.cornell.edu/preservation/IMLS>). The CDS model will provide centralized control for monitoring and managing distributed resources in all formats. The project report covers selection and content considerations, legal considerations, technical requirements for conversion, pre-deposit requirements, and metadata requirements. A Digital Preservation Officer (DPO) position was established in January 2002 to coordinate the development and implementation of preservation policy and serve as the liaison to digital preservation initiatives and projects. All new digital collection projects will be included within the scope of the CDS program and subject to its requirements.

The final component required by the OAIS model, the Dissemination Information Package, will address the needs of two distinct user groups. One group includes end-users. Both CUL and SUB have over a decade of experience in building access and navigation systems for digital collections, and through more recent work with Project Euclid, the user requirements of professional mathematicians and statisticians are increasingly understood. The other user group will be system administrators, who will require a secure interface into the system. In this work, recent redesign efforts of the administrative functions of the Physics arXiv by CUL will prove informative (see section 5.1.)

2.2 Preservation Metadata Development

Building a digital preservation and dissemination system based on the OAIS model calls for the development of metadata elements sets to represent the Content Information and Preservation Description Information associated with the digital objects preserved in the system. Content Information comprises the digital object itself and the Representation Information necessary to deliver and describe the object to future users, while Preservation Description Information retains the information needed to manage the object and its representation information over time. Though CUL and SUB recognize the importance of Content Information as a component of the Archival Information Package, much work has already been done in developing descriptive metadata, so they hold that the area of greatest need lies in the creation and evaluation of functional Preservation Description Information element sets. This need is underscored by the centrality of Preservation Description Information to the ongoing processes necessary to sustain an operational digital preservation system. Along with other groups working in the area of digital preservation, CUL and SUB recognize that "the elucidation and maintenance of Preservation Description Information is the keystone to building an information infrastructure to support the processes associated with digital preservation" [p. 46, OCLC, 2002]. CUL and SUB have designed the metadata components of the current project to address the need for and importance of preservation metadata element sets that meet the requirements of a functional digital preservation system.

Developing and assembling the metadata element sets to yield the Content and Preservation Description Information for the current project requires bridging the gap between the metadata already maintained by the content providers involved and those metadata actually needed to manage the digital preservation system. Moving from existing metadata to required metadata calls for two phases of work. The first involves surveying the content providers for the metadata now in use and building a map among the elements used with an eye toward future interoperability. The second phase involves a detailed analysis of OAIS documentation to determine those preservation description elements actually needed for long-term preservation and access. These undertakings will shape the foundation of the pre-ingest processing required to yield the Submission Information Packages to be added to the system. The project's metadata development work will also generate the structure for the system's Archival Information Packages.

The metadata development work associated with this project will generate preservation metadata element sets that CUL, SUB, and future collaborators can evaluate over time. More importantly, CUL and SUB will make those element sets available to the digital preservation communities at large for their review and refinement. This synergistic process can thus play a key role in the preservation metadata development that is so crucial at this stage in the history of digital preservation.

2.3 Electronic Mathematics Archives Network Initiative (EMANI)

In 2001, four academic libraries came together with two content providers to discuss the long-term preservation of mathematics literature in digital format. The project they initiated is the Electronic Mathematics Archives Network Initiative (EMANI). The four libraries are Cornell University Library, Göttingen State and University Library, Tsinghua University Library (Beijing), and Orsay Mathematical Library (Paris), and the content providers are Springer Verlag and the Electronic Library of The European Mathematical Information Service (EMIS). The project's mission is to develop and promote models for the preservation of digital content in mathematics. After an initial meeting in Boston in August 2001, a second meeting was held in Heidelberg in February 2002, to discuss the problems faced, possible solutions, and the focus of the project. Several broad principles were agreed to:

- the resulting archive should be “light;” that is, content should be accessible as soon after publication as is feasible
- the archive will not be directly revenue-generating
- archiving solutions should be developed collaboratively and be as widely accepted as possible, encouraging broad implementation and participation.

A second meeting was held in Göttingen in May, 2002, for preliminary discussions of metadata and file format issues related to a digital archive of mathematics literature.

Both CUL and SUB Göttingen see EMANI as providing a supportive context to the project proposed in this application. Springer Verlag and EMIS have agreed to provide content for the project, and to do the necessary conversion work to normalize this content.

3.0 PROJECT OBJECTIVES

3.1 Using OAIS as a guiding model, define the requirements for an interoperable, distributed preservation and dissemination system.

OAIS provides a high-level, abstract model of the functional needs of a digital archiving system. The first objective of this project will be to articulate that model at a much more specific level so that it meets the needs of a defined community and set of data. Two system design assumptions in the proposed project are not addressed by OAIS: that the system be distributed and interoperable.

The partners believe these architectural requirements are essential for the success of a digital archive for two reasons. First, distributed systems for the long-term retention and management of digital content are necessary because no single provider can be expected to maintain and ensure access to the archived literature of a single discipline, much less all archived literature. Such an approach would be unworkable technically and financially. Cost effective solutions to digital archiving must share responsibilities of long-term maintenance across numerous stakeholders.

Second, it is unrealistic to assume that different institutions, with different needs, resources, and users, will adopt common repository systems, even if they agree on the functional requirements of digital archiving. Digital librarians and technologists have noted this tendency toward multiple systems for some time, recognizing that what is needed are mechanisms that allow heterogeneous systems to communicate with each other. This is the basis for a number of digital library initiatives in the area of interoperability, including the Open Archives Initiative, and the existing project between the partners, “A Distributed Digital Library of Mathematical Monographs” (see section 5.1). The partners believe that extending this type of interoperability into the area of digital preservation is the most promising direction to move.

3.2 Develop the system architecture and workflows that integrate distributed repositories into a single preservation and dissemination system, including the rationales and means for exchanging content and its associated metadata.

The proposed project builds on the interoperability work of the Open Archives Initiative and “A Distributed Digital Library of Mathematical Monographs,” by extending cross-repository communication into the arena of digital preservation data. It assumes that requirements can be agreed upon to achieve digital asset longevity, and that these requirements may legitimately be met in a variety of ways by any number of disparate systems. In this project we will agree on a method of communicating information about these requirements.

The OAI and the Math Monograph project have been successful in developing protocol requests that serve as this communication layer. Once we understand requirements for the system we can articulate a protocol for sharing necessary information and satisfying system needs. This protocol work will inform local system development, in that the requirements of the distributed system must be met, while at the same time allowing local flexibility in system selection and implementation.

Such a protocol will create a single preservation system out of two separate systems. With an administrative interface that operates via the protocol layer, the system can be managed cooperatively through multiple (though common) administrative points of service. The protocol will also make the system extensible, allowing additional nodes to join the archiving system in the future.

3.3 Define the metadata elements necessary for the preservation and dissemination system.

CUL and SUB will design, implement, and evaluate the components of Submission Information Packages for the system. This will involve a survey of existing metadata and content format types from content providers, including Springer-Verlag, The European Mathematical Information Service (EMIS), and Project Euclid. CUL and SUB will work with content providers to establish pre-ingest processes to serve as exemplars for future collaborations. Adequate structural metadata will be developed to handle the complex digital objects typical of mathematics and statistics journal literature, which rely heavily on the TeX typesetting language. The project will offer the submission structures used in the project to the digital preservation communities at large for further evaluation and improvement.

CUL and SUB will design, implement, and evaluate the metadata element sets necessary to construct Archival Information Packages for the system. The AIPs implemented will address the critical need to describe, structure, and administer the complex objects submitted via the ingest processes. CUL and SUB will use OAIS guidelines, end-user access requirements, and system administrator access requirements to determine the Content and Preservation Description Information stored in the system's AIPs. We expect this work to rely on and extend the metadata development done in these areas by OCLC and RLG [OCLC, 2002]. The project will offer the content and preservation information structures used in the project to the digital preservation communities at large for further evaluation and improvement.

3.4 Build a working prototype that implements the developed design requirements for a preservation and dissemination system.

A major goal of the project will be to build a prototype that implements the design requirements developed in earlier project work. The prototype will allow CUL and SUB to test design choices, evaluate system architecture, and make refinements where needed. For content for the prototype, we will select from the resources provided by the publishers already mentioned, in consultation with mathematicians from our two institutions. We will work to ensure that the archived resources are linked to available online reviews and descriptive metadata in *Mathematical Reviews*, *Zentralblatt MATH*, and the *Jahrbuch über die Fortschritte der Mathematik*.

4.0 BROADER IMPACTS OF THE PROJECT

Mathematics is an enabling science for a host of other disciplines from the physical sciences to life sciences, as well as economics and social sciences. By creating an open digital archive of serial mathematics literature we will make mathematics reliably accessible to these other

disciplines. In addition, with the promise of persistent access, smaller libraries with limited resources will be able to forego archiving and storing print literature and instead access mathematics literature on-line.

For libraries and others addressing the problems of digital preservation, the project's compilation and development of metadata sets that satisfy the requirements of an OAIS compliant digital archive will contribute to research efforts in this field. The metadata standards we develop will be useable in other disciplines. We believe there are more similarities than differences between a digital mathematics archive and those in other subject areas. Thus our digital mathematics archive will serve as a prototype for other disciplines. For many, the protocols and metadata standards we develop will provide a basis upon which they can develop their own archives interoperable with ours. The resulting infrastructure will be broadly implementable, cost-effective, flexible, and extensible. Libraries that adopt the approach outlined here will be able to invest more in expanding content for targeted communities rather than developing infrastructure.

Trusted digital archives are an implicit component of the commitment to safeguard the rights of citizens through digital government, of initiatives to provide equitable access to information for comprehensive curriculum development, and of any endeavor that requires ongoing access to reliable information in response to specific criteria. The implementation of the OAIS model we are proposing offers the potential for delivering information to users without regard to traditional economic, political, and technological boundaries. The work will either be directly or conceptually transferable to other projects that are working towards collaborative delivery of shared resources.

5.0 PREVIOUS DIGITAL LIBRARY EFFORTS OF THE PARTNERS

5.1 Cornell University Library

For well over a decade, Cornell University Library (CUL) has been a leader in digital library efforts. It has created and maintained more than a dozen major digital library collections across a wide range of formats and disciplines from the Making of America Project, through the Cornell University Geospatial Information Repository, to the Core Historical Collection of Agriculture. In the process of developing and maintaining these collections and services, CUL has contributed to the creation of best practices in areas such as text and image conversion. Awards received include the Scout Award for the digital math books collection and the USDA Secretary's 1999 Honor Award for the USDA-Cornell Economics and Statistics System, and the ACRL Excellence in Academic Research Libraries Award for 2002. Current major initiatives include the Mellon-funded Project Euclid, a scholarly communication initiative to enable independent mathematics and statistics journals to publish their issues effectively and efficiently on the web as part of an aggregation.

Other Cornell units, namely Computer Science's Digital Library Research Group and Cornell Information Technologies have also contributed substantially to the research and practice of digital libraries. The library has a strong history of partnering with these units to leverage their expertise and to balance the different perspectives that these different units bring to the table.

Digital preservation has been a CUL focus for several years, both in research and practice. A major cooperative effort with the Cornell Computer Science Department is Project PRISM, a four-year, \$2.2 million, DLI2 funded effort. The Library's research team is characterizing the nature of preservation risks for Web-based resources, developing a risk management methodology for establishing a preservation monitoring and evaluation program, and developing management tools and policies for virtual remote control [PRISM]. Project Prism was featured in a January 2002 D-Lib article [Kenney, 2002]. The resulting framework will form the basis for developing an ongoing comprehensive monitoring program that is scalable, extensible, and cost effective. Project Prism is a collaboration of uniquely skilled librarians, computer scientists, evaluation experts, and international testbed participants.

Other digital preservation research projects and initiatives within Cornell University Library include an E-Journal Archiving planning grant and the Library's Common Depository System program (discussed in Section 2.1). In 2001, Cornell received one of seven planning grants that The Andrew W. Mellon Foundation awarded to develop digital archives for electronic journals. Cornell's effort, Project Harvest, used agricultural literature as the test case for a subject-based digital archive approach (SBDA). The resulting model explores the potential of new perspectives to establish a sustainable funding model and flexible, multi-tiered access models for digital archives that maximize the benefits of aggregation, made possible in a subject-based approach [HARVEST].

In the area of Mathematics publishing, CUL also has a long history, and an increasingly active present role. In 1991, 576 mathematics monographs were digitized as part of a preservation project. The titles were carefully selected by library staff and reviewed by a faculty advisory committee with an eye to their mathematical significance. Since this was a preservation project many very worthwhile candidate titles were not digitized because microfilm or reprint editions secured their preservation status. Even though this collection was constrained in its selection of titles it has proven to be very popular. From the beginning there has been a steady demand from libraries and individuals for printed copies of these books. With no active marketing effort more than 1,000 volumes have been sold to over 200 customers. The free online viewer for these books has generated a high level of use and has helped win recognition for the content of the collection (<http://cdl.library.cornell.edu/math.html>).

This collection of books is the basis of Cornell's content contributed to a current collaborative NSF supported project between The University of Michigan Library, Cornell, and the Göttingen State and University Library (Award Number: 0085853). This project, titled "A Distributed Digital Library of Mathematical Monographs," is creating a distributed digital library of historical mathematical literature, enhancing already rich, standards-based digital library systems at each of the institutions with mechanisms for interoperability. By means of this system, the three institutions will build a combined "virtual" collection of nearly 2,000 volumes of mathematical literature from the 19th and early 20th century, as well as of dissertations and related materials. The major work of this project has been the protocol development needed to support an abstracted exchange of search query information, although this has necessitated exposing structural and file format information to protocol requests as well. By defining and implementing an effective level of interoperability, this project aids users by creating "meta-

repositories" and uniform access mechanisms for focused subject collections. As well, it provides an extensible model for how interoperability can work for functionality beyond search query exchange. Much of the same logic and mechanisms can be applied to other types of information exchange, such as that needed for the verification of digital preservation criteria.

Project Euclid (<http://projecteuclid.org>) is an effort by CUL to build an online mechanism that facilitates electronic publication of proprietary scholarly literature in theoretical and applied mathematics and statistics [Koltay, 2002]. Sponsored by a grant from The Andrew W. Mellon Foundation, Project Euclid allows participating journals to publish on the Web in a timely, effective, and affordable manner, thereby increasing the visibility of their content through a combined online presence. The initial technical basis for Project Euclid was the system used by NCSTRL (Networked Computer Science Technical Reference Library), originally developed by Cornell's Department of Computer Science. This core has been modified and extended significantly to accommodate access controls, subscription services, and editorial services and is now a full repository architecture and protocol for proprietary documents. Its open architecture allows for flexible and extensible functionality and services. For example, Project Euclid facilitates reference linking by means of an automated service that extracts and parses references and then seeks article identifiers from various databases. Additionally, Euclid is compliant with the Open Archives Initiative, allowing open harvesting of all article-level metadata from its repository.

In the Fall of 2001, the Physics e-print archive, arXiv (<http://www.arxiv.org/>) moved to Cornell University. Since then CUL has maintained, housed, and administered the arXiv. CUL has recently begun a process to redesign the user interface and revise the ingest administration and user registration processes.

5.2 Göttingen State and University Library

Göttingen State and University Library (SUB) has been and is involved in more than 20 digital library projects. SUB holds twenty DFG-funded special collections (SSGs) in a variety of disciplines. These SSGs have been transformed into subject based Virtual Libraries with extended service functions like embedded search engines and portals. Resource discovery in a global environment is therefore one of the key issues for the daily work in Göttingen. Beginning with "Quality controlled Subject Gateways" for several subject fields and especially for Mathematics (MathGuide, <http://www.mathguide.de>), the library is now on the way toward developing a coordinated network for virtual subject based digital libraries including partners from Germany, Europe, and USA (libraries, databases, research institutions, publishers, etc.). Significant projects at SUB are progressing on the national, European, and international level in order to serve the mathematical community better with regard to digital resources.

SUB is now offering and maintaining two "one-stop-shops:" EULER (European Libraries and Electronic Resources in Mathematical Sciences) and RENARDUS (the academic Subject Gateway Service in Europe). EULER offers access to relevant mathematical resources like bibliographic databases (OPACs), Zentralblatt, Preprint servers, etc. RENARDUS offers access to all relevant quality controlled subject gateways so that mathematicians can search and browse for applications of Mathematics in other sciences (engineering, social sciences, etc.). EULER

and Renardus are designed in similar ways, both using a distributed architecture based on Z39.50 technology.

Göttingen has digitized a number of collections with significant activities in the fields of historical travel literature, North Americana and especially in Mathematics. The Jahrbuch-project, building up an Electronic Research Archive for Mathematics (ERAM), is a joint effort of Göttingen and the European Mathematical Society (Prof. Wegner). With DIEPER (DIgitised European PERiodicals) service, Göttingen Library leads a consortium of eight European Libraries, testing decentralized scanning-production and unified access over local repositories. Another feature of DIEPER is the European database for digitized documents, which like the EROMM (European Register of Microform Masters), serves as a central reference to avoid duplicate digital conversion. The database is located at Göttingen.

To achieve interoperability, Dublin Core based metadata elements play a central role in all these projects. Therefore, several members of the digital library project team at SUB are in close cooperation with the DCMI initiative, especially with the DC Advisory Board (one of the team is a member of the board). The project results feed into some important Dublin Core Working Groups (DC Library, DC Registry, DC Type).

Within the program frame of establishing a distributed digital research library in Germany, a new program has been initiated to support retrospective digitization of library holdings. The Göttingen Digitization Center (GDZ), established in May 1997, is one of two national supply centers for digitization in Germany, with the second center located at the Bavarian State Library in Munich. The focus of the activities at the GDZ has been on the different fields of technology required to build a digital library. Following the recommendations of the DFG, the GDZ chose a strategy of collaboration with an industrial software partner to create a Document Management System (DMS) as a key component for the digital library. Agora, the new DMS, was presented in Göttingen in April 1999 and is now in production at the GDZ. In order to allow for maximum interoperability with other metadata sources, the system works with an import/export format that is based on RDF and XML. Advanced searches in metadata for different document types (e.g., monographs, multi-volume works, and journals) can be combined with searching in document structures such as chapters, articles, and figures. It is also possible to browse single or multiple collections. Agora developers integrated the Verity Information Server, a powerful full-text search engine, used in a number of significant digital library efforts. The inclusion of Verity now makes it possible to offer effective searching, in metadata such as bibliographic fields, titles of chapters, articles, and figures. Göttingen is preparing to offer full-text searching as well, a feature that becomes increasingly important as Göttingen moves from digitization of older text material (often in *Fraktur* type) to 20th century works. Verity is able to search a wide range of document formats from MS-Word over PDF to XML files.

Recently the SUB Göttingen has joined several digital library projects, e.g. the MathDiss International project (DFG funded), the ProPrint project (DFN funded), an NSF/DFG project together with Oldenburg University (Germany) and Virginia Tech (USA) which pertains to extended OAI applications.

5.3 The Partnership—A Collaborative History

The two institutions have a great deal in common, including their rich collections in mathematics, their strong digital library efforts, their interest in the problems of access (especially with regard to multi-lingual collections of a single discipline) and their involvement in national and international collaboration (including with each other).

The establishment of the Digitization Center (GDZ) at Göttingen State and University Library is closely connected to Cornell University Library. During the establishment of the Center in May 1997, Norbert Lossau, head of the GDZ, together with Frank Klaproth, visited a number of libraries in the U.S. with a focus on digital library activities. Cornell was of special interest for Göttingen because of their extensive experience in fundamental research in the field of digitization techniques and their success in organizing the production of the digital conversion process. Subsequently, in order to share U.S. experiences with a larger audience of German librarians, Anne R. Kenney from Cornell was invited to the first national workshop of the two German Digitization Centers (Göttingen, January 1998). Later, for the third German workshop (Göttingen, October 1999), Sandra Payette from the Cornell Department of Computer Sciences's Digital Library Research Group discussed their successful efforts at Cornell. As part of the continuing exchange with their US colleagues, Norbert Lossau again visited Cornell in November 1999. During that visit and subsequently, the operations at Göttingen and Cornell have continued to exchange information and experience regarding procedures for sustaining a high quality of metadata capture and with regard to techniques to make digital documents available via the WWW. At Cornell, Norbert Lossau gave a lecture about "Document Management for digitized Books: RDF/XML as solution for mirroring complex metadata- and document structures," and participated in intensive discussions about features of the Göttingen Agora and various Cornell delivery systems. Beyond these personal contacts, the GDZ remains in close communication with Cornell with regard to various topics of digital library research and production. In 2001 a further project has been started, named "THE DEVELOPMENT OF A DISTRIBUTED DIGITAL LIBRARY OF MATHEMATICAL MONOGRAPHS". Several meetings took place in Göttingen and Cornell. The last meeting was held at SUB in May 2002.

Since last year, both institutions have been active participants in another international project: the EMANI (Electronic Mathematical Archiving Network Initiative) collaboration, which began at the IFLA meeting in Boston, August 2001. Since then, two meetings (February 2002 in Heidelberg at Springer Verlag and in July 2002 at Cornell University) were organized. As a special bilateral activity in the context of EMANI, SUB hosted in May 2002 a metadata workshop with strong participation of Cornell.

Both institutions feel strongly that the proposed cooperative activity is the result of significant ongoing dialogue and should prove of great value for digital library efforts in Germany and the U.S.

6.0 PLAN OF WORK

Phase 1: Staff at Göttingen and Cornell will define the functional requirements of a preservation system by bringing the OAIS elements to an operational level, by building on past experience in developing retrieval systems, and by working to develop a shared vision of the needs of user groups, such as end users and staff users. This will include decisions regarding the file formats to be preserved and the normalization of source files. One of the institutions will take the lead, preparing a draft for circulation. Deadlines for input will be established with the goal of focusing the work of project meetings to the most complex intellectual tasks.

Phase 2: Concurrent with Phase 1, staff will survey content providers, developing an inventory of file and object types typical for in mathematics journal literature. An understanding of file management needs will be developed in light of the common use of the TeX typesetting language for mathematics literature. The question of file normalization will be addressed. Again, preliminary work will be carried out initially by one partner, who will then pass reports to the other partner for review and comment. In this way, there will be an exchange of work deadlines in preparation for project meetings.

Phase 3: Concurrent with Phase 1 and 2, staff will inventory preservation metadata sets in use at partner sites, at content provider sites, at other digital preservation implementations, and as defined by the digital preservation communities. The metadata areas in need of development for Submission Information Packages and Archival Information Packages will be identified and some preliminary element design work will be done independently by the partners and circulated. Project meetings will be held to establish the metadata structures necessary for SIPs and AIPs. Staff will develop and document the pre-ingest processes required for content providers and project staff to assemble SIPs.

Phase 4: Staff will develop tools to capture and enrich metadata for the system. This work will be based upon the results of Phases 1-3.

Phase 5: Staff will develop a system architecture, metadata framework, and protocol for building the interoperability layer that joins partner repositories. In this phase, staff will analyze the functional requirements identified in Phase 1 in order to isolate aspects of the system that depend on the interoperability layer. The interoperability layer will handle all exchanges between nodes in the distributed archival system. Detailed protocol specifications will be developed or enhanced to support the required interoperability functionality. Considerations include architecture implementation decisions that may facilitate participation by future partners via a shared-system/API model.

Phase 6: Staff will modify or enhance existing local systems to achieve system architecture interoperability requirements. The first step in this phase will be to map the proposed system architecture to local system architectures in order to determine which functional components exist locally, which need modification, and which need to be developed. A development plan will then be drawn up which details the work needed to complete the necessary functional components at the local level. We will then modify local repository frameworks to support new object types and preservation metadata sets. At this point, local systems should support the

functional requirements of the proposed archival system. The final step in this phase involves implementing the interoperability protocol developed in phase 5. This will allow us to begin to exchange information between distributed archival servers and to test the interoperability layer between partners.

Phase 7: Staff will populate the system with selected content. Content will be a selection of resources from participating publishers. These include representatives from three groups typical of the publishing scene in mathematics: professional societies (EMIS - The European Mathematical Information Service), large commercial publishers (Springer-Verlag), and small, independent publishers and distributors (Cornell's Project Euclid).

Phase 8: Staff will implement access and management interfaces sufficient to the purpose of evaluating the system.

Phase 9: Staff will use evaluation data to refine infrastructure components (workflow, sustainability, communication, metadata element sets, etc.) necessary to develop a full production system.

Phase 10: Concurrent with all phases, staff will disseminate results of their work, via papers and presentations at appropriate venues, such as Digital Library Federation forums, Library and Information Technology Associations working groups and meetings, *D-Lib Magazine*, *RLG DigiNews*, and other appropriate conventions, meetings, and publications.

7.0 MANAGEMENT PLAN

There will be three Co-Principal investigators of the project. At CUL, Marcy Rosenkrantz, Director of Library Systems and H. Thomas Hickerson, Associate University Librarian for Information Technologies and Special Collections and Director of the Division of Digital Library and Information Technologies, will be Co-Principal Investigators. Prof. Dr. Elmar Mittler, Director of SUB will be the third PI.

Overall management of the project will be under the direction of the Project Director, David Ruddy. He will develop the detailed work plan and ensure that deadlines are met. Heike Neuroth will be project manager at SUB and share responsibility with David Ruddy for coordinating the work between the two institutions.