

International Paper Registry Technical Proposal and Work Plan Trefoil Corporation July 15, 2008

Key Personnel:

Curtis Meadow, Senior Software Engineer, Trefoil Corporation. Responsible for database design and programming, application design and programming, Bernstein integration

Marilyn Lutz, Director, Information Technology Planning, Raymond H. Fogler Library, University of Maine. Responsible for authority file research, design of authority holding structures for names, titles and subject headings and general conformance with the current best practices of library cataloging for special collections; OAI profile; research on unique identifiers.

Graphic Designer, TBA

Task 1. Analysis and modification of database structure (Months 1-3)

A review of the current IPR database structure shows that it is based on a well-designed third normal form physical data model. The physical data model derives from a logical entity-relationship model that is based on extensive analysis of the problem domain of describing paper substrates. Hence the basic structure is not expected to require significant modification. However, development of a new implementation of the database to function in a distributed database environment requires analysis, including evaluation of alternative approaches, and modifications for certain items, as follows:

A. Establish system of unique record identifiers needed for use via a distributed database system

Due to the fungible nature of much of the IPR domain there are few "natural" primary keys amongst the entities in the domain. Most records are therefore identified by arbitrary primary keys. Ensuring unique primary keys is a requirement for a distributed database.

We plan to use Universally Unique Identifiers (UUIDs) to generate primary keys. UUIDs however are not "user-friendly." They are formatted hexadecimal numbers 34 characters in length, for example "{fb63089e-852f-4e46-99ca-36e852d09968}." They must be hidden from the user, even if we have a to generate a more user-friendly pseudo-key such as a sequential integer.

B. Add coded lookup tables for controlled vocabularies to allow multi-lingual search and output from the database. (Parts of Tasks 4 and 5)

1. Wherever possible, we will use standardized vocabularies used in bibliographic cataloging. Examples of items that can be converted are Artist - Role (use MARC/Dublin Core relator terms) and Artist - Language (Use ISO 639-2 or ISO 3166-1 language codes). We will identify usable existing standard vocabularies in this first task of the work plan.
2. Determine optimal structure for lookup tables and coding schema for coded values. Because coding in the form of simple numeric values introduce the possibility of duplication at different sites in the distributed database system, we will use values stored in a canonical language (English).

C. Establish XML database structure mappings to allow exchange of data between nodes of the distributed database with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH is an XML protocol based on the Dublin Core Metadata Initiative (DCMI) that provides for exchange of data between repositories. Publication of OAI data will also allow libraries and other

institutions to obtain paper data from the database.

D. Establish a date storage format. Review date storage formats to determine whether the specification in the current IPR prototype should be fully adopted because of its adaptability to dating systems of different cultures and historical eras, or if a standard such as ISO 8601 should be adopted. Date storage format has some implications for database management system portability, since the syntax for querying dates varies among platforms.

E. Review available authority files and determine which will be used and where.

Names and Subject headings are available from OCLC, the Library of Congress (LC) and the Consortium of European Libraries (CERL). The bibliographic record is treated as the authority for uniform titles; titles of works are available for SRU or Z39.50 lookup from OCLC, LC and the European Library, as well as many other sources. The Getty Foundation publishes the Art & Architecture Thesaurus (AAT), the Union List of Artist Names (ULAN) and the Getty Thesaurus of Geographic Names (TGN)

F. Modify data structures where necessary to accommodate authority file references.

G. Add metadata tables for labeling tables and columns for multi-lingual output.

H. Analyze security and audit trail requirements for data entry, search and retrieval, and design suitable structures and/or database permissions. Security is a significant requirement if the IPR or any of its nodes will be used for document authentication and detection of forgeries.

I. Add "record suppression" flags to appropriate table(s) so that individual records can be suppressed from public search and display.

Deliverables:

1. Documents detailing design decisions and the rationale for the decisions.
2. Complete Entity-Relationship (ER) and physical design models. Physical design models will be in Boyce-Codd Normal Form (BCNF). If empirical testing indicates that denormalized views are necessary or useful they will be specified.

Total Estimated Time:

Lutz: 140 hours

Meadow: 160 hours

Other Budget Items:

We are not budgeting for OCLC subscription costs, because during development we will operate under the existing Bates College OCLC subscription. LC offers automated title searches at no cost using either SRU or Z39.50 protocols, and OCLC Research currently offers a web service that provides LC authority file lookup at no cost. Getty Thesauri are licensed to non-profits at \$750 per thesaurus for a 5-year license; we are budgeting \$2,250 towards this expense.

Task 2. Creation of an empty prototype relational structure in MySQL. (Months 3-4)

This task includes establishing referential integrity constraints, populating controlled vocabulary lookup tables, populating metadata tables and establishing check constraints. Although the structure will be created and defined in MySQL, it will be independent of the back end database, allowing it to be easily

ported by means of one or more Data Definition Language (DDL) scripts to other relational databases such as Oracle, SQL Server, SQL Anywhere, Microsoft Access.

Deliverables:

1. Documentation of constraints and controlled vocabularies
2. DDL database creation scripts for MySQL and a selection of other popular relational databases.

Total Estimated Time:

Meadow: 40 hours

Task 3. Build the data entry front-end (Months 4-10)

The data entry front-end will be constructed using open-source software. PHP appears to be the best choice for the server scripting language as it is open-source and is available for Windows, Macintosh and Unix/Linux. To the extent possible, the code will be independent of both the web server and the database server. If some dependencies must exist, they will be isolated and well-documented.

Some parts of the public search-and-display front-end will be incorporated in the data entry front end. In particular, searches are needed for the data entry side as well as the public side. The basic record display should be functional in the data entry side as well as the public side.

Subtasks:

1. Workflow analysis. We will work with Prof. Allison to determine normal workflow for cataloging typical items. Well-designed screens for data entry should be based on the workflow rather than the physical structure of a database. Although navigation between screens is subject to relational and other integrity constraints, the screens should allow data to be entered in a natural sequence without a need to memorize complex rules.
2. Graphic design. A good graphic design is important for end-users. A good graphic design will enhance performance (efficiency) of users by intuitive relationship to work flow, simplicity of work screens. Poor graphic design can inhibit dissemination of the end-products by intimidating or daunting prospective users, especially since many users will not be paper experts.

Guidelines for graphic design:

- A. Presentation will be controlled by Cascading Style Sheets (CSS) version 2, thereby freeing server scripting to deal solely with content.
 - B. Graphic design will be clean and simple, making minimal use of graphic elements.
 - C. Relative positioning ("flow layout") will be used in preference to absolute positioning.
 - D. Graphic design will be compliant with Section 508 of the Americans with Disabilities Act (ADA)
 - E. The target browser code platform is XHTML 1.0 Transitional
3. Screen and Navigation Design. Initial templates will be pure HTML with CSS. These will be produced by the graphic designer and approved by the project before being implemented in server-side scripting.
 4. Server-side scripting and testing. Scripting will be compatible with a secure sockets layer (SSL) connection. Because it is possible that institutions installing a copy of the open source system may not have a server certificate, the data entry scripts will be designed to operate over

either a secure HTTPS connection or a standard HTTP connection. Connection type will be configurable by a switch in a configuration preferences file.

5. Data entry search and browse. Search and browse features for the data entry front end are needed to retrieve records for viewing and editing. Search and browse will be handled directly by SQL queries.

Deliverables:

1. Workflow analysis document
2. Graphic Design templates
3. Data entry website

Total Estimated time/Costs

Meadow	400 hours	
Lutz	40 hours	
Graphic Designer	TBA	\$5,000

Task 4. Create / Configure SRU Gateway (Months 10-12)

Search/Retrieval via URL (SRU) is an XML-focused search protocol for Internet search queries, using CQL (Contextual Query Language), derived from the Z39.50 specifications. An SRU (or SRW) gateway takes a CQL query as input; processes the query and submits the query to the database server in its native language (SQL in this case) and then processes the result set and returns the results formatted in XML.

We will adapt Bernstein code or other open-source SRU implementations (for example, British Library) to the IPR.

Total Estimated time/Costs

Meadow	120 hours	
--------	-----------	--

Task 5. Year 2 Enhancements (Months 13-24)

Year 2 will be devoted to continued testing, debugging and user interface enhancements derived from user feedback. It is well-known that the number of bugs discovered in software decreases exponentially over time. A project of this size may involve 30,000 to 60,000 lines of code, between PHP scripting, database queries and stored procedures in SQL, client-side code in JavaScript and HTML and CSS used in PHP templates. We expect the first month after completion of the data entry side to be moderately busy with testing and debugging, with little or no activity needed by the completion of the second year.

In addition to debugging, we expect that some end-user enhancements will be requested. This is the normal course of software development – abstract specifications are never a complete substitute for sitting down at a keyboard and performing a task.

Total Estimated Time/Costs

Meadow	160 hours	
Lutz	20 hours	

Task 6. Documentation and Dissemination (Distribution) (Months 15-21)

Documentation will consist of both internal documentation (listed above in deliverables sections) and documentation for other users of the IPR. The latter documentation will be largely the responsibility of Allison and Hart. We expect to contribute some technical sections to this document.

We propose to make the IPR database available for distribution as a website with an accompanying MySQL database. In this case "website" means a set of files, organized in a set of folders, that can be copied onto a machine and configured within a web server as a standalone website. The distribution will include:

1. Complete data entry website (PHP, HTML, CSS files and image files)
2. Documentation for configuration under common web servers (e.g., Apache, IIS)
3. An empty MySQL Database
4. SQL Data Definition Language (DDL) scripts and documentation for creating an empty database in MySQL and other popular database management systems (e.g., SQL Server, Oracle, Microsoft Access)
5. SRU gateway with configuration instructions
6. If applicable, a public website with configuration documentation

Total Estimated Time/Costs

Meadow	80 hours
Lutz	20 hours