ECHO DEPository - Phase 2: 2008-2010 Final Report of Project Activities

Beth Sandore and John Unsworth, Principal Investigators
University of Illinois at Urbana-Champaign
June 30, 2010

PROJECT TEAM MEMBERS

Hyoungtae Cho, Matt Cordial, Sarah Dotson, David Dubin, Janet Eke, Joseph Futrelle, Elizabeth German, C. Jean Godby, Thomas Habing, Myung-Ja Han, Mamta Hsing, Patricia Hswe, William Ingram, Larry S. Jackson, Judith Klavans, Rebecca LaPlante, Robert Manaster, Jerome McDonough, Lev Ratinov, Dan Roth, Carolyn Sheffield, Devon Smith, and Guojun Zhu.

Table of Contents

1. Executive Summary

Phase 1 Activities

In Phase 1 (2004-2007) of the ECHO DEPository (ECHO DEP) Project, the University of Illinois at Urbana-Champaign (UIUC) partnered with the Online Computer Library Center (OCLC); the National Center for Supercomputing Applications (NCSA), based at UIUC; Michigan State University Libraries; and an alliance of state libraries in Arizona, Connecticut, Illinois, North Carolina, and Wisconsin to work on a set of technical architecture projects addressing the challenges of digital preservation.

A special issue of Library Trends (Cruse, P. & Sandore, B.; eds., 2009) comprises sixteen articles that tell fascinating stories about the ground-breaking efforts of numerous partners within the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP). Since its inception in 2004, NDIIPP has grown from an experimental program into a true partnership of concerned organizations working together to sustain access to digital information that is critical to scholarship and cultural heritage nationwide. Each article in this issue tells a compelling story conveying the sense of urgency that has pervaded the efforts of the numerous institutions and groups involved in NDIIPP, many with little else in common but the need to develop policy, structure, process, commitments, and technologies to preserve significant cultural and historical content into the future.

Phase 1 activities focused on:

- developing the Web Archives Workbench tool, in partnership with OCLC and several state libraries and archives. The Web Archives Workbench is a suite of web capture tools based on principles of managing archived content in aggregates rather than as individual objects. The suite comprises:
    1. The Discovery Tool, which helps identify potentially relevant websites by crawling relevant “seed” entry points to generate a list of domains to which they link;
    2. The Properties Tool, which enables users to maintain information about content creators, associate them with the websites they are responsible for, and enter high-level metadata;
    3. The Analysis Tool, which permits user to look at the structure of the website to see what kind of content is represented by the file directory; and
    4. The Harvest Tool, which allows user to monitor crawl status, review and modify harvest settings, and package harvests for transfer to a repository. The Harvest Tools also offers a separate Quick Harvest feature that

schedules one-time harvests of content. Harvest packages are encoded in METS with Dublin Core metadata embedded.

- a systematic evaluation of existing repository software applications. In our repository evaluation, we tested the suitability of several open-source repositories to ingest and manage different types of content (encoded texts, html documents, temporal media). We also evaluated the stability of these repositories in handling a high volume of digital objects. The results of this work, which were shared with the community in both technical reports and published articles, provided useful information for institutions choosing repository software in connection with digital preservation efforts.
- designing an architecture that enhances the interoperability and preservation features present in repository software. As the repository evaluation team's work got off the ground they discovered that they needed to design an environment that enabled the capture and retention of full preservation metadata to support the transfer of objects and metadata from one repository to another without loss of critical information and functionality. The interplay between digital preservation metadata was critical to the team's design and development of workflows and systems to ingest and manage digital content (for example, in the METS profile work that our NDIIPP team did with LOC and other institutions). Other institutions have now adopted the Hub and Spoke model for migrating digital content across multiple repositories.
- modeling next-generation repositories for long-term preservation. In Phase 1, the Semantic Archive (SemArch) team of the Graduate School of Library and Information Science (GSLIS) at UIUC and NCSA approached problem of semantic preservation by exploring how to infer meaning from digital object structures that change over time. The outcome of this work was a model of semantic inference that should help next-generation archives head off long-term preservation risks. Phase 2 then developed a model and proof-of-concept implementation for the reliable inference of provenance metadata about a digital object. To carry out this work, SemArch
  1. scoped and assessed various content types
  2. constructed generalized ontologies for harmonizing with grid provenance work
  3. explored and analyzed the Open Provenance Model; and
  4. experimented with intentional query processing strategies for efficient interrogation of remote resource collections.

  The semantic preservation component of our project contributed in at least two ways to the community's understanding of interactions between the digital preservation process and the meaning carried in what was being preserved:

1. It provided a better understanding of the (sometimes unanticipated) changes to structural and semantic properties of digital content that can be introduced in the process of archiving and migration; and
2. It helped us to establish that not all existing repository packages ingest metadata and object information equally well, and that this can cause information loss at the very beginning of the digital preservation process.

Phase 2 Activities

The University of Illinois and OCLC were further funded under the NDIIPP program, with new partners at the University of Maryland and the Department of Computer Science at UIUC, to build upon several Phase 1 deliverables in a Phase 2 effort (2007-2010). Phase 2 continued our research and development in repository interoperability and semantic preservation, and launched two new technical architecture projects: tools for metadata creation and extraction, and tools for risk assessment of data formats across long-term preservation.

- The Hub and Spoke (HandS) repository architecture of Phase 1, which supported the management and transfer of content among repository systems such as DSpace, Eprints, and Fedora, was extended. After considering commercial applications, including CONTENTdm, LOCKSS (Lots of Copies Keep Stuff Safe) and JSR 170 (which eliminates the need for learning proprietary APIs by providing a standard one), SWORD (Simple Web-service Offering Repository Deposit) was implemented as the fourth spoke. A "Master METS" profile was added to the repository architecture, acting as a manifest of various other "snapshot METS" profiles.
- The Extracting Metadata for Preservation (EMP) project, new to Phase 2, developed a generalized metadata tool architecture and built a documented, open-source tool for high-quality named-entity metadata extraction and creation, including mapping to authority files. The tool supports named-entity metadata extraction from both existing structured marked-up text (metadata extraction) and from free text (metadata creation). Our approach was based on providing machine assistance for human creation of metadata, using linguistic technology. EMP built on work occurring at OCLC, at the UIUC Department of Computer Science, and at the University of Maryland (UMD). Much of the project was spent evaluating named-entity extractor tools, a process which helped determine the gold standard used in evaluating the EMP tool. In addition, EMP explored in detail what integration of its named-entity extractor tool with UMD's CliMB (Computational Linguistics for Metadata Building) toolkit would involve, as a proof-of-concept exercise for future phases of the project.

- The INFORM project, also new to Phase 2, built an online infrastructure, implementing the INFORM methodology for assessment of data format risk, setting up test groups to apply the methodology, then analyzing the results and incorporating them into the Global Data Format Registry (GDFR). Media preservation librarians and other information professionals working with media formats were selected to test the tool.

2. Introduction

2.1 About the ECHO DEPository Project – Phase 2

In Phase 1 (2004-2007) of the ECHO DEPository (ECHO DEP) Project, the University of Illinois at Urbana-Champaign (UIUC) partnered with the following institutions to work on a set of technical architecture projects addressing the challenges of digital preservation: the Online Computer Library Center (OCLC); the National Center for Supercomputing Applications (NCSA), based at UIUC; Michigan State University Libraries; and an alliance of state libraries in Arizona, Connecticut, Illinois, North Carolina, and Wisconsin. Phase 1 activities focused on the development of web archiving tools; a systematic evaluation of existing repository software applications; the development of an architecture to enhance interoperability and preservation features present in repository software; and the modeling of next-generation repositories for supporting long-term preservation. Project participants reported on these activities in depth in the ECHO DEPository Phase 1 Final Report, submitted to the Library of Congress in July 2009.

At the end of 2007 the University of Illinois was approved for further funding under the NDIIPP program to build upon several Phase 1 deliverables. In January 2008 Phase 2 of ECHO DEP was launched. During this phase we continued research and development in the areas of repository interoperability and semantic preservation and also launched two new technical architecture projects, namely tools for metadata creation and extraction and for preservation risk assessment of data formats. While OCLC continued as a project partner in Phase 2 of ECHO DEP, we also gained two new partners during this period: the University of Maryland and the Department of Computer Science at UIUC.

2.2 About this Document

This narrative report provides an overview of ECHO DEP Phase 2 activities and accomplishments. Accompanying appendices and referenced download facilities make available specific additional project deliverables (e.g., EMP's evaluation of existing NER tools and the ontology developed by the Semantic Archive project). Many of these materials, as well as this report itself, are archived in IDEALS[1], the institutional repository at the University of Illinois, and thus available for future public access.

A note about nomenclature: to reduce confusion about phases (e.g., there is a Phase 2 of ECHO DEP, but there are also phases of development within a discrete

[1] http://www.ideals.illinois.edu/

project), the second phase of ECHO DEP will herein be termed "ECHO DEP 2." (Accordingly, "ECHO DEP 1" denotes the first phase of ECHO DEP.)

2.3 Review of Project Objectives and Deliverables

2.3.1 Hub and Spoke - HandS (repository interoperability project)

Goals

- Develop open-source interoperability tools – specifically, expanding the Hub and Spoke (HandS) architecture to include additional open-source or commercial repositories.
- Build community collaboration, including collaboration to define the characteristics of a preferred AIP.
- Build additional format-specific METS sub-profiles.

Deliverables

- Ongoing maintenance of current functionality, such as:
    - Improvements to the code base, toward a fully-stable, production-quality release.
    - Updates to the latest relevant versions of currently utilized toolkits (XML beans; JHOVE).
    - Updates to new versions of Dspace.
    - Maintenance of our METS profile.
- Development of new HandS functionality
    - Implementation of additional popular repositories.
    - Expanding of framework to make use of new services and file formats.
- Potential deliverable: Expand HandS Suite to make use of Global Digital Format Registry (dependent on GDFR implementation)

Overview

A key outcome of our repository interoperability activities during Phase 1 of ECHO DEP was the HandS repository architecture, which supports the management and transfer of content among repository systems, such as DSpace, Eprints, and Fedora. During Phase 2 we researched and discussed the issue of adding another "spoke" (repository software application) to our framework. After considering commercial applications, including CONTENTdm, and other possibilities, such as LOCKSS (Lots of Copies Keep Stuff Safe)[2] and JSR 170 (which eliminates the need for learning proprietary APIs by providing a standard one), the HandS team decided to implement SWORD (Simple Web-service Offering Repository Deposit)[3] as the fourth spoke. In addition, we continued work on our METS profile by introducing a "Master METS" profile into our repository architecture – a profile that acts as a manifest of various other "snapshot METS" profiles.

Details

Please see Section 3 of this report.

2.3.2 Semantic archive – SemArch (semantic preservation project)

Goal

- Development of a rules-based inference model and proof-of-concept implementation for identifying preservation risks and inferring provenance information from content collections.

Deliverables

- Development of a model for rules-based automatic generation of provenance metadata.
- Development of a proof-of-concept implementation.

Overview

During Phase 1 of ECHO DEP, faculty at the Graduate School of Library and Information Science (GSLIS) collaborated with research programmers at NCSA on the problem of semantic preservation. Most repository systems preserve the structure of a digital object; few, if any, however, preserve its *meaning,* or semantics. The Semantic Archive (SemArch) team approached this problem by exploring how to infer meaning from digital object structures that change over time. The outcome of this work was a model of semantic inference capability to help next-generation archives head off long-term preservation risks.

In ECHO DEP Phase 2, SemArch essentially continued along this track. It developed a model and proof-of-concept implementation for the reliable inference of provenance metadata about a digital object. To carry out this work, SemArch scoped and assessed content types; constructed generalized ontologies for harmonizing with grid provenance work; explored and analyzed the Open

[2] http://www.lockss.org/lockss/Home

[3] http://www.swordapp.org/

Provenance Model[4]; and experimented with intentional query processing strategies for efficient interrogation of remote resource collections (determined as a result of scalability analysis).

Details

Please see Section 4 of this report.

2.3.3 Extracting metadata for preservation - EMP (named entity recognition and extraction project)

Goals

- To develop a generalized metadata tool architecture.
- To build a named entity metadata extraction tool.
- Development will be based on two approaches:
    - Extracting from existing structured marked-up text (i.e., metadata extraction).
    - Extracting from free text, or "buckets of words" (i.e., metadata creation).

Deliverables

- Development of a documented open-source tool for high quality named entity metadata extraction and creation, including mapping to authority files.
    - Approaches to support extraction may include development of external metadata profiles or development of a machine-learning approach.
- Development of a general metadata tool architecture extensible to other types of metadata tools.
- An evaluation and analysis of existing named entity metadata tools.

Overview

Unlike HandS and SemArch, which continued projects begun in Phase 1 of ECHO DEP, the Extracting Metadata for Preservation (EMP) project started its development activities in Phase 2. Building on work occurring at OCLC, at the UIUC Department of Computer Science, and at the University of Maryland (UMD), the EMP project developed stand-alone, open-source tools for automated metadata extraction.

[4] http://openprovenance.org/

With digital content ever increasing, there is a need for improving the efficiency of metadata creation and extraction. One approach is to provide machine-assistance for the creation of metadata by using linguistic technology. Specifically, EMP proposed to develop a generalized metadata tool architecture and build a Named Entity Metadata Extraction Tool. A good measure of the project was spent evaluating named entity extractor tools, a process which helped us determine the gold standard for our own developing tool. In addition, we explored in detail what integration of our named entity extractor tool with UMD's CliMB (Computational Linguistics for Metadata Building) toolkit[5] would involve, as a proof-of-concept exercise for future phases of the project.

Details

Please see Section 5 of this report.

2.3.4 INFORM (INFORM risk assessment methodology implementation project)

Goals

The goal of the INFORM project is to build an infrastructure to support a community-developed information resource for assessing data format risk. This includes:

- Implementing the methodology as an online application.
- Building analysis software.
- Building test groups to apply methodology, analyzing results and incorporating them into the GDFR.

Deliverables

The above goals involved three main areas of activity, each corresponding to a deliverable:

1. **Software development:** Implement the INFORM methodology as a software application.
2. **Research protocol development:** Design the research protocol; build a community of experts to apply the methodology.
3. **Data-gathering and analysis:** Gather data from participants; analyze and review.

[5] http://www.umiacs.umd.edu/~climb/

Overview

The INFORM Risk Assessment Methodology Project was another project that began in Phase 2. It addressed the uncertainty surrounding the curation of data formats by building a collaborative environment for assessing data format risk.[6] The methodology behind the INFORM tool defines risk categories of digital formats, as well as the risk factors for each category. It also scales to measure probability of occurrence and impact. Media preservation librarians and other information professionals working with media formats tested the INFORM assessment tool, and this report relates our findings.

Details

Please see Section 6 of this report.

3. Hub and Spoke (HandS) Project

In continuing the HandS project, work was carried out in three phases:

- **Phase 1** – Ongoing maintenance of the tool suite
- **Phase 2** – New development, such as implementation of new spokes
- **Phase 3** – Development wrap-up and software releases

An understanding of HandS activities during Phase 2 of ECHO DEP depends in part on prior knowledge of what took place during Phase 1 of the project. Thus, a brief summary of HandS development during Phase 1 of ECHO DEP is included here (see Section 3.1). For a complete, detailed representation of HandS activities carried out during Phase 1, please refer to the *ECHO DEPository Technical Architecture Phase 1 Final Report* (pp. 36-57).

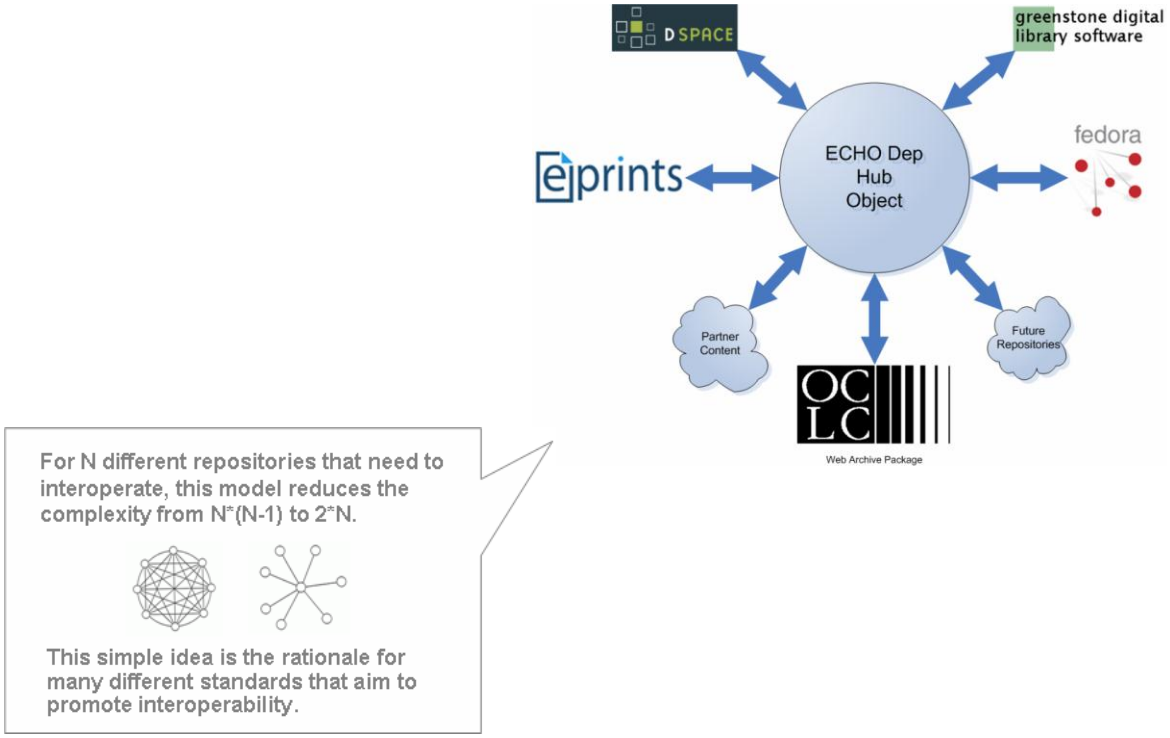
3.1 Summary of HandS Activities during Phase 1 of ECHO DEP

The HandS architecture grew out of activities at the start of ECHO DEP 1, in which team members evaluated a range of open-source repository software applications. A key finding of the evaluation was inefficient support in these applications for interoperability, as well as minimal adherence to preservation standards. Current institutional storage practices also warrant effective tools for repository interoperability: often, more than one repository software application is deployed at an institution, thus making necessary the transfer and management of content between installations. Finally, rarely is repository software static, or unchanging;

[6] More about the INFORM approach is found here: http://www.dlib.org/dlib/november04/stanescu/11stanescu.html.

new versions of a platform are a given, making interoperability a fundamental requirement for preservation of digital objects.

These findings spurred the development of the HandS suite of tools, created to help libraries manage and preserve content in a multiple repository setting (Habing, Eke, Cordial, Ingram, Manaster, 2009). They also enable safekeeping of valuable preservation data. HandS tools apply a common standards-based method (a PREMIS-based METS profile) for packaging content, allowing digital objects to be transferred between repositories easily while facilitating the collecting of technical and provenance information imperative for long-term preservation. This model (simplified in Figure 1 below) has been implemented in several real-world archiving projects.



For N different repositories that need to interoperate, this model reduces the complexity from N*(N-1) to 2*N.

This simple idea is the rationale for many different standards that aim to promote interoperability.

Figure 1: A graphic representation of the HandS framework.
For a functional overview and details on workflow, see the ECHO DEPository Technical Architecture Phase 1 Final Report (pp. 36-57).

3.2 Phase 1 – Ongoing Maintenance of the Tool Suite

3.2.1 Producing a stable, production-quality release

Because of the small project and staff size, ongoing development and maintenance of the software followed an informal process. As they were discovered, bugs and issues with the code were discussed and prioritized during regular project staff meetings, where they were assigned to a programmer to be fixed. Bugs were mostly discovered informally as part of the normal development and deployment process. However, formal testing did occur as part of our large-scale ingestion tests. During these tests the output of the OCLC Web Archives Workbench (WAW) tool, developed during ECHO DEP 1 (2004-2007), was submitted and then disseminated to and from various supported repositories using the Hub and Spoke system. These tests did help discover several software problems that were fixed during this phase.

The primary tool used to track changes to the code was the Subversion version control system hosted at Sourceforge. All code changes were submitted to this system along with brief commit log messages describing the changes being committed. Late in 2008 an “echodep-commits” mailing list was established.[7] All commits to the Subversion database generate an email to subscribers of this list. The commits mailing list allowed project staff or other interested parties to more easily track changes to the code.

During ECHO DEP 2 there have been four public release packages:

1. Version 0.5, 2008-01-30, represented the code at the end of Phase I
2. Version 0.5.1, 2008-02-27, was a minor bug fix release
3. Version 0.6, 2009-04-14, was the first release of Phase II code
4. Version 0.8, 2010-01-22, is the final release of code for Phase II.

The following sections describe in more detail some of the significant changes to the software during this phase.

3.2.2 DSpace

In spring 2008 DSpace released version 1.5, a major refactoring of its codebase. Portions of our Lightweight Repository Create/Retrieve/Update/Delete (LRCRUD) service for DSpace had to be rewritten in order to work with the new version. However, we anticipated that many institutions currently running the earlier version of DSpace would not want to upgrade (at least not right away). Changing our LRCRUD would prohibit these institutions from using our service on their older DSpace installations. Instead, we decided to create a new spoke for DSpace 1.5, and leave intact the older spoke for DSpace 1.4. Doing so revealed an added benefit: it

[7] http://sourceforge.net/mailarchive/forum.php?forum_name=echodep-commits

was now possible to use the Hub and Spoke for migrating packages from older versions of DSpace to the new version, DSpace 1.5.

3.2.3 Eprints

In the early spring of 2009, we upgraded to EPrints to EPrints-3.1.2.1 (Chocolate-coated Coffee Bean). It was, at the time, the latest stable version. At present, the latest stable version is 3.1.3. Besides being modular and flexible enough to work within the Hub and Spoke framework (and its improvements, such as the addition of other repository spokes), this version provided us with support for the Simple Web-service Offering Repository Deposit, or SWORD. (For more explanation of the role that SWORD played in HandS activities, please see Section 3.3.2.) Changes in the EPrints code were made to accommodate our revised HandS model using the master METS file. The LRCRUD code, which is written in Perl, also was updated to work with the upgraded EPrints. Finally, we arranged the EPrints LRCRUD documentation (and program code) more effectively, so that it could be packaged in a Java and remain relatively independent of the rest of the Hub and Spoke architecture. In essence, the EPrints spoke gave us a proof of concept: it confirmed that a spoke could be written, documented, and packaged in a language and platform independent from the environment in which the hub and the other spokes were written (Java). This relative independence shows the power and flexibility of not only the Hub and Spoke model but of any implementation of this model as well.

3.2.4 METS profiles

Restructuring

Probably the most significant change to the Hub and Spoke system resulted from a restructuring of our METS profiles. These changes were initiated toward the end of ECHO DEP 1, but they were finished during ECHO DEP 2. The original design, which was implemented during ECHO DEP 1, required the instantiation of a single METS file into which all the changes made to the underlying preservation entities were recorded. Although not implemented during ECHO DEP 1, we also explored concepts related to tracking and merging metadata changes into a single primary metadata section in this single METS file. After some initial exploration we realized that doing this within a single file would result in large, complex, and unwieldy files, and that trying to simplify the resulting tangle by merging changes into new metadata sections was an intractable problem (at least within the context of our project) and would likely result in data losses, in any case. Because of these complications, we decided instead to look at the concept of versioning—that is, a versioning of entire METS files, where each file would represent a different version or snapshot of the METS package at a point in time. This approach significantly simplified the individual snapshot METS files, while at the same time preserving the history of all changes in previous versions of the file. However, with this change we now needed a way to track and ensure the preservation of all the different versions of the METS files associated with the preservation of the given entity. For this purpose we designed a "Master METS Profile" (also known as "master METS file"). The master METS file consists of a single `<structMap>` containing multiple div elements each with an `<mptr>` element that points to one of the snapshot METS files. To enhance their preservation, the Master METS file also contains PREMIS preservation metadata for each of the snapshot files. The following image (Figure 2) illustrates a master METS file, which references three snapshot METS files which in turn reference multiple files making up the entity to be preserved.

The addition of the Master METS file required significant changes to our underlying code, which have been completed as part of ECHO DEP 2. With these changes to our code, any significant preservation activities—namely, the submission or dissemination to, or from, a repository—result in a new snapshot METS file and the addition of that snapshot file to the Master METS file. The new Master METS profile has been documented and registered with the Library of Congress.[8]

Maintenance

Other than publishing the new Master METS profile, no maintenance was required for the METS profiles originally published as part of ECHO DEP 1. Several of the XML Schema used by the METS profiles underwent revisions during the course of the project. METS went from version 1.6 to 1.8; MODS from version 3.2 to 3.3; and PREMIS from version 1.1 to 2.0. Other technical metadata schema such as MIX also underwent revisions during this period, but since they were not as critical as the other three schema to the METS profile, they were not as carefully tracked.

The METS and MODS revisions were minor enough that documents conforming to our profile could still be validated using the newer XML schema, and none of the changes to these schema would have substantially changed the functionality, or form, of our current profile. Therefore, we did not make any changes to our profiles or underlying code to accommodate these new schema.

Although the changes to the PREMIS XML schema and the PREMIS data dictionary from version 1.1 to 2.0 were substantial and not backward compatible, we felt that the usage of PREMIS in our profile would not be substantially altered by these changes. In reality, our profile's requirements for PREMIS are minimal and can easily be accommodated by either version of the schema. However, this unfortunately means that the more expressive capabilities of the new PREMIS Schema will not be available to users of our METS Profiles. If there is any future work on these profiles, changes to accommodate PREMIS 2.0 should be a priority.

[8] http://www.loc.gov/mets/profiles/00000029.xml.

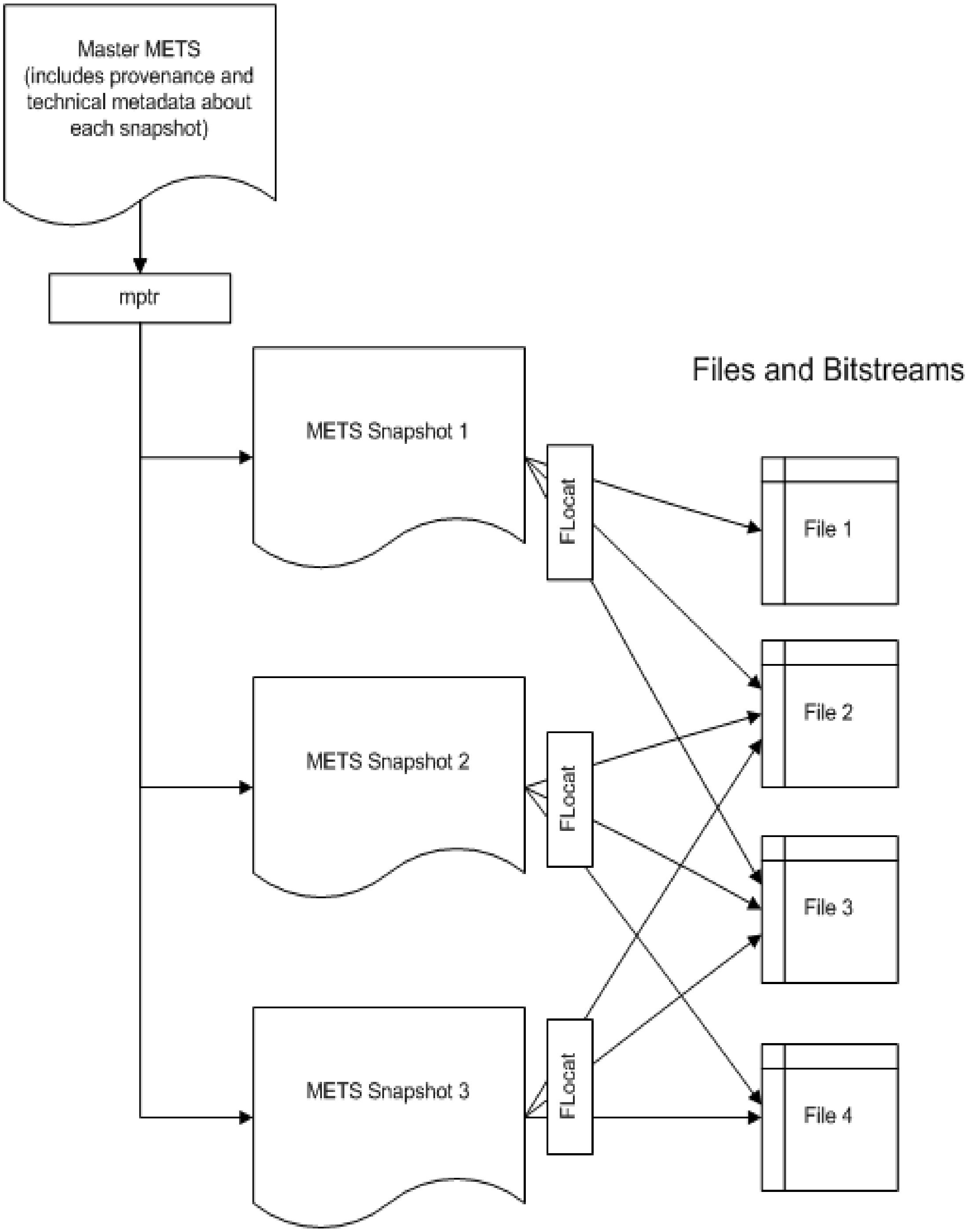


Figure 2. Illustration of a Master METS file, referencing three snapshot METS files, which in turn reference the multiple files that make up the entity bring preserved.

Standards participation

During the course of the project, staff members participated in several functions related to the METS and PREMIS metadata standards. These included attendance at METS editorial board meetings as schedules allowed (mostly during the DLF Forum post-conference meetings), presentations at several METS and PREMIS tutorial or conference sessions hosted by LoC, and participation in an ongoing work group developing best practices for using PREMIS in METS.

3.3 Phase 2 - New Development

Besides maintaining the existing code, a significant amount of work during this phase went toward adding new functionality to the Hub and Spoke system. This new functionality was mostly focused on adding support for new repositories, like Fedora, or providing support for standard exchange protocols like SWORD or BagIt. It also went toward usability improvements, such as providing a graphical user interface for basic package transfers between repositories. In addition, planned improvements such as integration with the GDFR or JHOVE II were not realized, either because these systems were not ready within the time-frame of this project, or because of unanticipated changes to the organizational responsibility for them. Sections 3.3.1 through 3.3.5 below provide additional detail on each of these new developments.

3.3.1 Fedora

Development of the Fedora spoke included Hub-to-Fedora and Fedora-to-Hub Packager modules, facilitating interoperability through pluggable interfaces, and an LRCRUD service for Fedora, supporting the dissemination and submission of objects by defining a protocol for transmitting digital objects to and from a Fedora repository over HTTP.

Fedora packagers

Development of the Hub-to-Fedora and Fedora-to-Hub Packager modules was based on our previously developed DSpace packagers. The Hub-to-Fedora packager creates Fedora Object XML (FOXML) from the HandS METS files. These METS files are contained along with the package content files in the Fedora Object. The FOXML, HandS METS, and content files are then copied into a ZIP file for transport to the repository. The Fedora-to-Hub packager takes the native Fedora dissemination, unpacks it, and creates new HandS METS Profile objects from the contents.

Fedora LRCRUD

The LRCRUD Service for Fedora sits alongside the repository and provides a Representational State Transfer (REST) interface for Fedora. (CRUD is an acronym for Create Retrieve Update and Delete - implemented as the HTTP methods POST, GET, PUT, and DELETE, respectively). The implementation of the Fedora LRCUD service was loosely based on our LRCRUD service for DSpace, with an important exception concerning the way items are ingested into the repository. Unlike DSpace, which accepts a package of files for ingest, Fedora requires that all content files are referenced from the FOXML via URLs. In other words, all the content files to be ingested into Fedora will first need to be made available on the Web. Only the FOXML for the package is ingested directly into the repository; the rest of the package is retrieved by the repository itself, from URLs indicated in the FOXML.

This requirement of Web-available files presented a challenge. Our solution was to expose a temporary staging directory to the Web and allow the Fedora LRCURD service to unzip the package contents into the directory. In order for this to work, the Hub-to-Fedora Packager module would have to be able to retrieve the location of this staging directory so it could correctly indicate URLs for the content files in the FOXML. When the Fedora LRCURD receives a package for ingestion, it unzips the package contents to the staging directory and ingests the FOXML, and once Fedora has retrieved all the package contents, they are deleted by the LRCRUD service from the staging directory.

3.3.2 SWORD

In fall 2008 the Hub and Spoke team began considering various additional repository software applications to incorporate another “spoke” in the Hub and Spoke architecture. Up to this point, the spokes that were supported in the framework were DSpace, Eprints, and OCLC Digital Archive—with activity underway to add Fedora. Under consideration were the CONTENTdm[9], especially since the University of Illinois Libraries have many digital collections supported by the Cdm; Simple Web-service Offering Repository Deposit (SWORD)[10]; Lots of Copies Keep Stuff Safe (LOCKSS)[11]; and JSR 170[12]. In the end, SWORD was selected as the final repository software application, since momentum had begun to build behind it in the repository interoperability community (and there was interest in SWORD’s Microsoft Word plug-in, enabling Word documents to be deposited directly into a repository). In addition, the latest versions of DSpace, Fedora, and Eprints—which were either already spokes, or would soon be a spoke, in the Hub and Spoke architecture—all support SWORD.

A lightweight protocol for depositing content into a repository, SWORD has two parts to the deposit process:

[9] http://www.contentdm.org/

[10] http://www.swordapp.org/

[11] http://www.lockss.org/lockss/Home

[12] http://www.day.com/content/dam/day/whitepapers/JSR_170_White_Paper.pdf

1. A request is made to the repository from an authenticated user for what is called a "service document". Depending on the user's credentials, the service document provides a list of collections the user is allowed to deposit into.
2. SWORD-enabled repositories provide special deposit URLs for sending items directly into collections listed in the service document without any further interaction with the repository itself. If the deposited item is a recognized package of data and metadata, the packaged metadata will be used for describing the package to the repository automatically.

The SWORD technology is much like our LRCRUD service, except that it can only be used for deposit.

SWORD packager

In order for an information package to be deposited into a SWORD-enabled repository, it must be packaged in a format that the repository recognizes. The repositories in our test bed—Dspace, Fedora, and Eprints—accept DSpace, METS-based submission packages with embedded metadata in the Scholarly Works Application Profile (SWAP) format. SWAP is a Dublin Core Application Profile for describing scholarly works. The SWAP model is based on the Functional Requirements for Bibliographic Records (FRBR) idea of dividing metadata elements into types: Work, Expression, Manifestation, and Item.

By far our biggest challenge to creating "SWORD packages" - as we call them - was to crosswalk the Hub package MODS metadata into SWAP. With the help of a UIUC metadata librarian, we developed a MODS-to-SWAP XSLT stylesheet for handling the transformations.[13] The SWORD Packager generates the SWAP metadata and embeds it in a simple METS file, which serves as the manifest for the package. All the content files and metadata, including our EchoDep METS files, are included. Once the package is complete, it can be sent to any SWORD-enabled repository.

3.3.3 BagIt

In July 2008 the Library of Congress announced its support for the BagIt format.[14] They intended to use this format, along with associated exchange protocols and scripts, to transfer content—generated by partners during the first phase of NDIIPP—to the Library of Congress for archiving.

BagIt support was not in the original workplan for Hub and Spoke. However, since the University of Illinois had content from ECHO DEP 1 that needed to be

[13] For a description of the challenges we encountered in crosswalking from MODS to SWAP, please see slides 46-53 of our presentation, "Repository Interoperability and Preservation: The Hub and Spoke Framework," delivered at the DLF Spring 2009 Forum: http://www.diglib.org/forums/spring2009/presentations/Habing.pdf.

[14] http://www.cdlib.org/inside/diglib/bagit/bagitspec.html

transferred, and that content was already in the Hub and Spoke METS package format, it was relatively easy to add a customized export spoke for BagIt to our Hub and Spoke system.

Because of the simplicity of the BagIt format, this addition of support was fairly easy with the first implementation committed to our source code repository in August 2008. The only real technical challenge was efficiently generating the checksums. Using our BagIt spoke we generated about 5000 packages, which could range from one file to hundreds of files per package; the total extent came to approximately 16 GB of files. A bag of bags was created containing all the individual packages. This bag was posted to our website, from whence the Library of Congress downloaded it.

Project staff provided formal presentations on our BagIt implementation and content transfer experience at several meetings and conferences including the DLF Fall Forum 2008 and the NDIIPP Partners meeting June, 2009. In addition, our BagIt implementation proved useful for other NDIIPP projects at Illinois, namely the Preserving Virtual Worlds project[15] for which we customized a Spoke for packaging a directory structure containing game data into a Hub and Spoke package that was then converted to a BagIt package.

3.3.4 Graphical user interface

Previous versions of the HandS client tools could be used only via a command line interface (CLI). With the new SWORD spoke, however, came the need for a more interactive user interface. A SWORD deposit requires three-steps:

1. retrieve a service document
2. select a deposit target from the service document
3. deposit an item

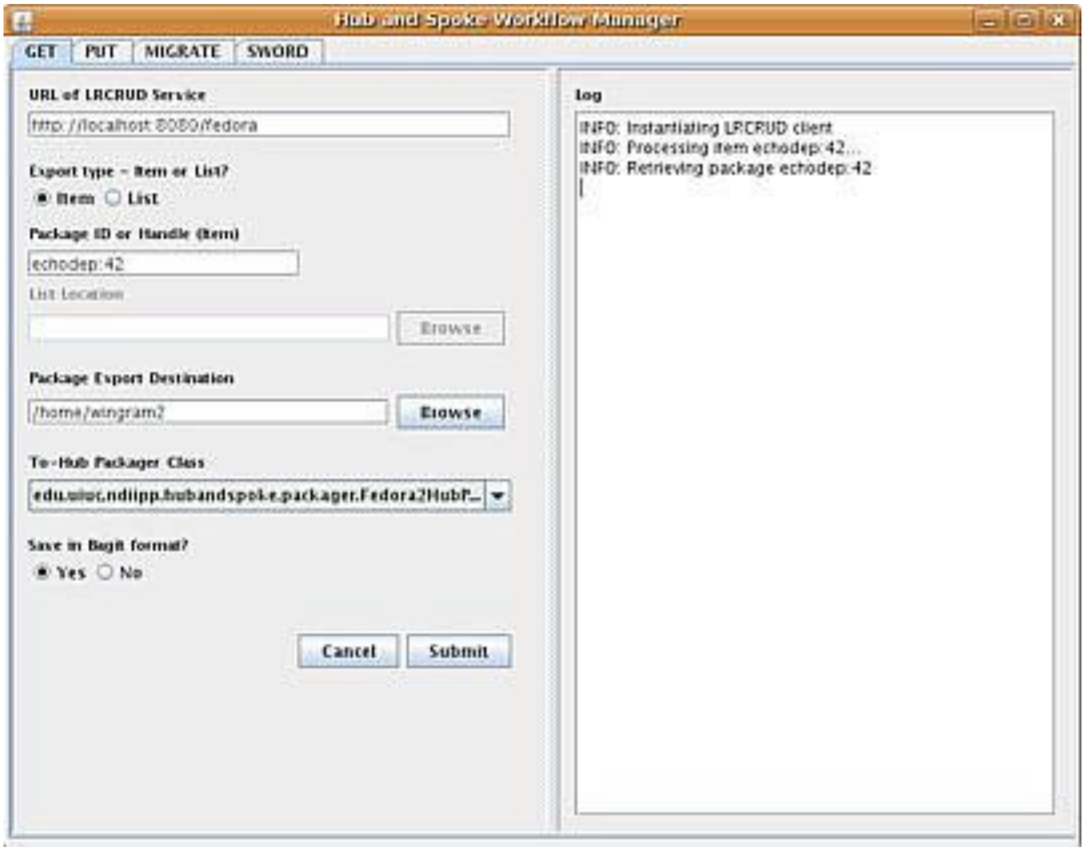
We found that a graphical user interface (GUI) lends itself to this process better than our old CLI. We designed the new GUI to facilitate the most common uses of the HandS client tools, with modules for interacting with SWORD-enabled repositories and for depositing, retrieving, and migrating packages to and from repositories via the LRCRUD protocol.

The GET module

The GET module is for retrieving items from a repository via LRCRUD, as shown in Figure 3. To retrieve an item, the user enters the URL of the LRCRUD service for the repository containing the desired item(s). The module can be set to retrieve a single item given its repository-item ID, or handle, or multiple items listed in a text file. The user sets the export destination, and chooses a To-Hub Packager from a drop-

[15] http://pvw.illinois.edu/pvw/

down menu, which is automatically populated with the names of the installed packager modules. Finally, the user may select the option to have the retrieved package converted into the BagIt format for archiving. The log window shows the progress of the package retrieval, and displays any warnings or error messages that may have occurred.



URL of LRCRUD Service
http://localhost:8080/fedora

Export type – Item or List?
● Item ○ List

Package ID or Handle (Item)
echodep:42

List Location

Package Export Destination
/home/wingram2

To-Hub Packager Class
edu.uiuc.ndiipp.hubandspoke.packager.Fedora2HubP...

Save in BagIt format?
● Yes ○ No

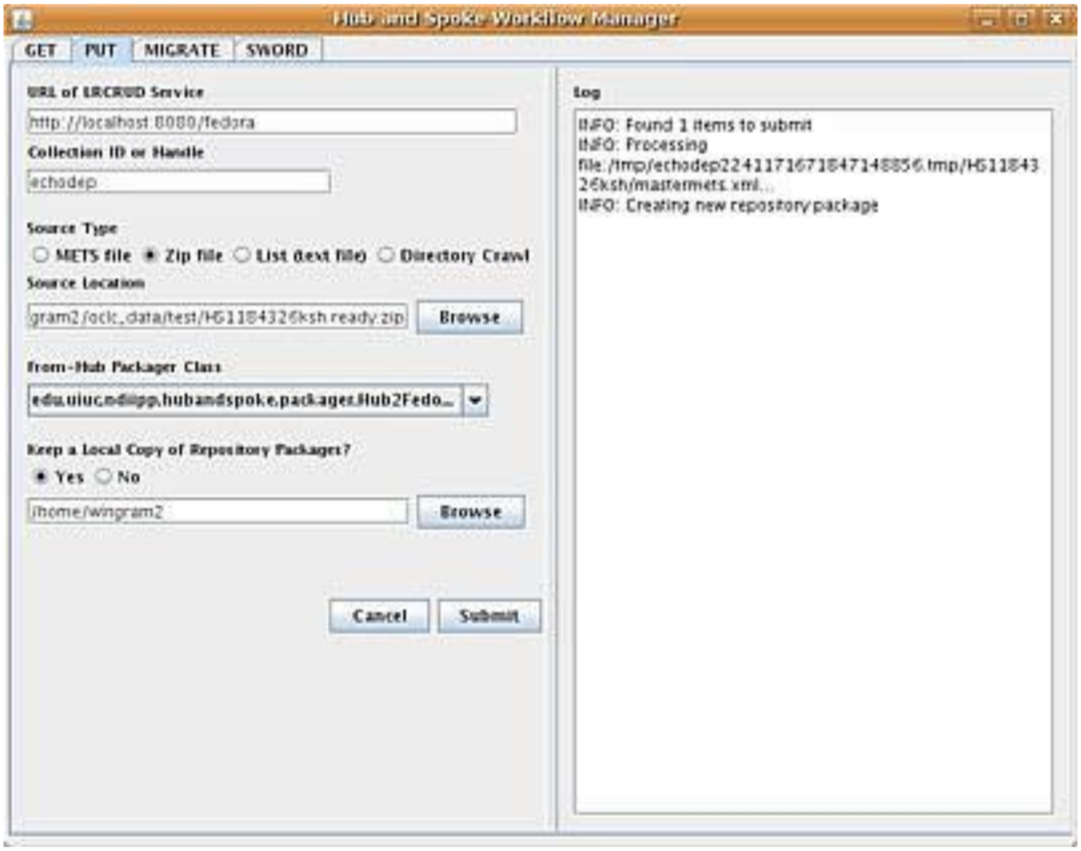
Log

```
INFO: Instantiating LRCRUD client
INFO: Processing item echodep:42...
INFO: Retrieving package echodep:42
```

Figure 3. The GET module in the GUI for the Hub and Spoke Workflow Manager.

The PUT module

The PUT module is for depositing items in a repository via LRCRUD, as shown in Figure 4. To deposit an item, the user enters the URL of the LRCRUD service for the target repository, and the target collection ID or namespace. The source type can be set to METS file, ZIP file, List (text file), or Directory Crawl. The first two source types refer to a single item - either to an EchoDep METS file for the package or to the package itself, if zipped. The other two source types can be a list of packages (METS or ZIP), or directory containing one or more packages. The From-Hub packager class must be set to the packager for the target repository. If the user wishes to keep a copy of the packaged item, a location directory must be designated. The log window shows the progress of the packaging and depositing, warnings or error messages, and for successful deposits it displays the repository location (or handle) of the deposited item.



URL of LRCRUD Service
http://localhost:8080/fedora

Collection ID or Handle
echodep

Source Type
○ METS file ● Zip file ○ List (text file) ○ Directory Crawl

Source Location
gram2/oclc_data/test/HS1184326ksh.ready.zip

From-Hub Packager Class
edu.uiuc.ndiipp.hubandspoke.packager.Hub2Fedo...

Keep a Local Copy of Repository Packages?
● Yes ○ No
/home/wingram2

Log

```
INFO: Found 1 items to submit
INFO: Processing
file:/tmp/echodep2241171671847148856.tmp/HS11843
26ksh/mastermets.xml...
INFO: Creating new repository package
```

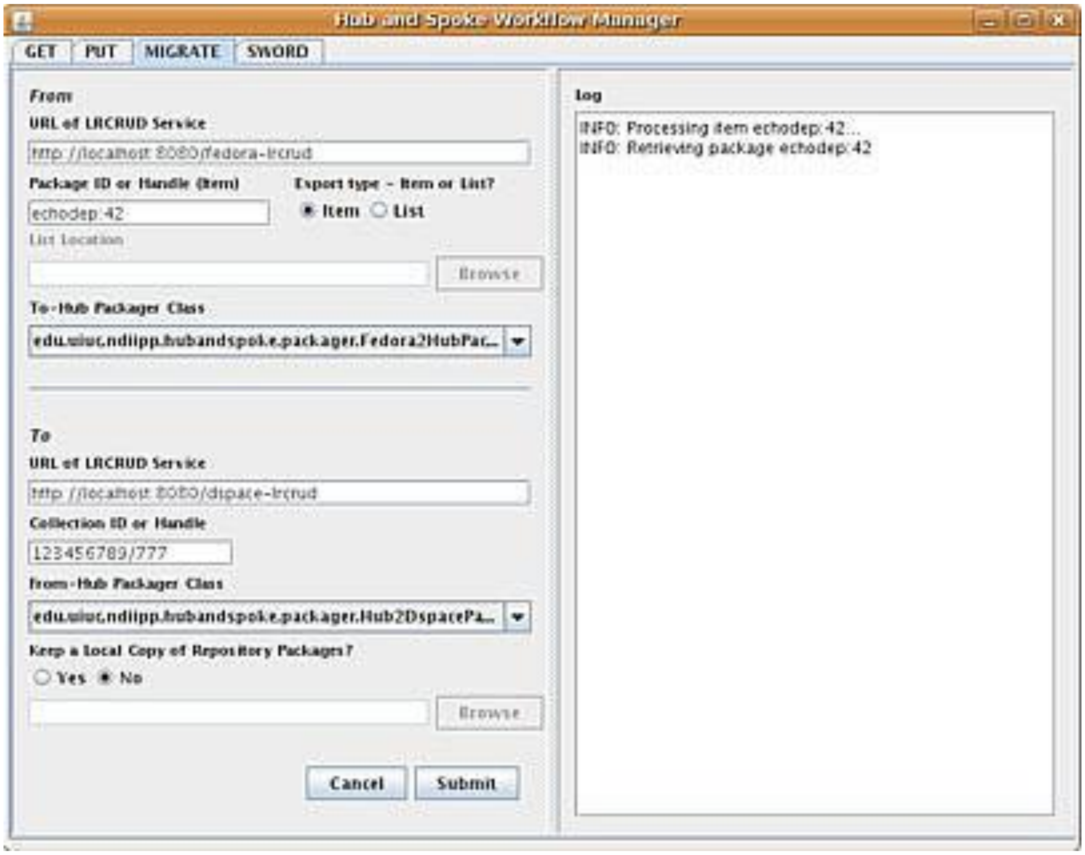
Figure 4. The PUT module in the GUI for the workflow manager.

The MIGRATE module

The MIGRATE module, shown in Figure 5, is used for conveniently copying an item between repositories, or between locations in the same repository. It simply combines the GET and PUT modules.

The SWORD module

The SWORD module, displayed in Figure 6, retrieves SWORD service documents and makes deposits into SWORD-enabled repositories.



From

URL of LRCRUD Service: http://localhost:8080/fedora-lrcrud

Package ID or Handle (Item): echodep:42

Export type – Item or List? ● Item ○ List

List Location:

To-Hub Packager Class: edu.uiuc.ndiipp.hubandspoke.packager.Fedora2HubPac...

To

URL of LRCRUD Service: http://localhost:8080/dspace-lrcrud

Collection ID or Handle: 123456789/777

From-Hub Packager Class: edu.uiuc.ndiipp.hubandspoke.packager.Hub2DspacePa...

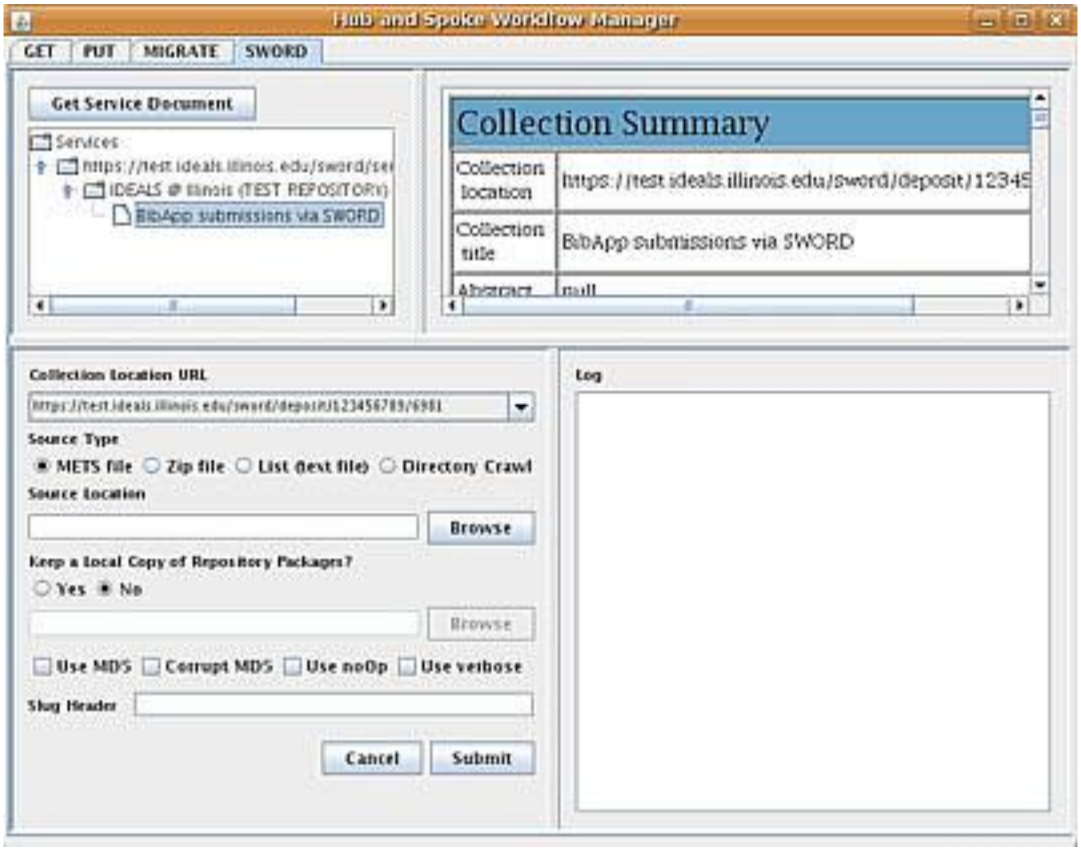
Keep a Local Copy of Repository Packages? ○ Yes ● No

Log

```
INFO: Processing item echodep:42...
INFO: Retrieving package echodep:42
```

Figure 5. The MIGRATE module in the GUI for the workflow manager.

The area on top is used to retrieve the service document, and display metadata for the available collections, including any license agreements or restrictions. Once a target collection has been selected, the item may be deposited using the interface below the service document area. These deposit options are very similar to those of the PUT module, with additional SWORD settings that are used mainly for testing and verifying the status of the deposit location. The module also allows users to enter a "Slug header"; this allows repositories to support the Slug entity-header for naming the to-be-created repository location.



Services
https://test.ideals.illinois.edu/sword/se
IDEALS @ Illinois (TEST REPOSITORY)
BibApp submissions via SWORD

Collection Summary

Collection location	https://test.ideals.illinois.edu/sword/deposit/12345
Collection title	BibApp submissions via SWORD
Abstract	null

Collection Location URL
https://test.ideals.illinois.edu/sword/deposit/123456789/6981

Source Type
● METS file ○ Zip file ○ List (text file) ○ Directory Crawl

Source Location

Keep a Local Copy of Repository Packages?
○ Yes ● No

□ Use MD5 □ Corrupt MD5 □ Use noOp □ Use verbose

Slug Header

Figure 6. The SWORD module in the GUI for the workflow manager.

3.3.5 Global Digital Format Registry (GDFR)

A potential HandS deliverable when ECHO DEP 2 began was the incorporation of format information from the GDFR into METS profile technical metadata. However, in mid-2008, OCLC withdrew from the registry project (at that point, the GDFR began being hosted by Harvard), and a year later—at the 2009 NDIIPP Partners Meeting—it was announced that the GDFR would be merging with PRONOM (a U.K.-based format registry project) to form the Unified Digital Format Registry, or the UDFR. With the uncertainty brought on by these developments, the HandS project team decided against pursuing the task of integrating GDFR format information into our METS profile.

3.4 Phase 3 – Development Wrap-up and Software Releases

A significant portion of the last months of development was devoted to documentation and release packaging. For the documentation, a library graduate assistant with experience writing technical documentation was employed. The documentation is in the form of a website.[16] It is also included in our official software release package which is available from SourceForge.[17] The last version released as part of this project is version 0.8, 2010-01-22. The documentation includes installation instructions, JavaDocs for developers, and command line and graphical user interface usage instructions for end-users, as well as references to related material such as METS profiles, presentations, and published papers.

The source code itself is available in the release 0.8 TAR file from SourceForge, but may also be checked-out directly from the SourceForge Subversion source code repository. If we make any future changes to the code, we do plan to continue to post updates to SourceForge, but we have no immediate plans for this.

3.5 Conclusions and Open Questions

While there has been much progress in the digital preservation community since the inception of the HandS project in mid-2006, we still feel that repository interoperability is of major importance to long-term digital preservation, and there remains work to be done. We are grateful to the Library of Congress for extending our project, and we feel that the ECHO DEP 2 follow-on to the HandS project allowed us to validate our initial architectural design and implementation and to improve upon it.

The addition of new spokes, including ones for Fedora, SWORD, and BagIt, while improving the general usefulness of the system, has also allowed us to improve the modularity and extensibility of our software code which we hope will allow potential users to easily create their own pluggable packagers for addition repositories. We at Illinois will be utilizing this extensibility to support future work on our own preservation repository, and our hope is that others in the preservation community will find the application equally useful. The only high-level architectural change to our application was the addition of the Master METS concept. We feel that this, or a similar concept, is important for any preservation system which anticipates that digital objects will change over time, for example as the result of format migrations. The Master METS allows for simple, controlled versioning of preservation packages over time with support for validation and provenance tracking of the new packages. We also hope that the addition of a graphical user interface (GUI) to the system will make it more useful.

Like any development project, there are things that have been left undone or that we wish we could improve on. Among these is our LRCRUD protocol. Given that it was developed at about the same time as the Atom Publishing Protocol (APP) and

[16] http://dli.grainger.uiuc.edu/echodep/hands/index.html

[17] https://sourceforge.net/projects/echodep/

before the SWORD protocol (which is a profile of APP), it was fairly leading-edge. However, there are aspects of the Atom Publishing and SWORD protocols that are more flexible and powerful than our LRCRUD protocol, probably the most significant feature being the idea of Service Documents. Given additional time and resources, and considering that the APP is an IETF standard, we would have considered swapping LRCRUD for SWORD or some similar APP profile.

Another area where we would have like to spend additional development time was in our Fedora spoke. Given the constraints of our project, our Fedora spoke implementation is fairly simplistic. Digital objects that make up a package are ingested into Fedora in a flat structure, or when they are disseminated from Fedora they are converted to a flat structure. This is similar to how our DSpace packagers work, primarily because DSpace is only capable of managing flat lists of files. However, Fedora is better able to accommodate more complexly structured digital objects, using its RELS-EXT and RELS-INT datastreams to assert relationships between objects. Early in our design stages we explored how we might be able to utilize these Fedora capabilities to more accurately reflect complex structural maps for our METS packages being ingested into Fedora, or how these relationships could be represented in our METS packages when objects were disseminated from Fedora. We quickly realized that developing and implementing mappings that could accommodate arbitrary RELS-EXT and RELS-INT relationships in Fedora into METS structural maps and vice versa could be a complex and lengthy project all by itself. Given that our objective was primarily proof of concept, we opted for the simpler approach. However, we feel that it would still be worthwhile to explore the more complex approach, and for any digital object which requires its complex internal relationships to be accurately expressed after it is ingested into Fedora (or a similar repository) this mapping is critical.

4. Semantic Archive (SemArch) Project

Work on the SemArch project was carried out in three phases during ECHO DEP 2:

- **Phase 1:** Scoping and assessment
- **Phase 2:** Implementation
- **Phase 3:** Documentation and release

Activities focusing on the challenges of semantic preservation evolved from the repository interoperability work that occurred during ECHO DEP 1. Below is a review of the research that set the foundation for the SemArch project in ECHO DEP 2.

4.1 Summary of Semantic Preservation Research during ECHO DEP 1

Phase 1 of the ECHODEP project took us from case and scenario-specific analyses of preservation to a more general characterization of the limitations of traditional metadata description. We developed a framework for the role of inference in discovering specific examples of those weaknesses and targeting resources at risk (Dubin, Futrelle, and Plutchak, 2006). Specifically, we had:

- Reviewed the literature of digital object definitions
- Examined problems in mapping structural relationships among electronic records to logical dependencies (in the context of a legacy administrative database).
- **L**ooked at some issues of dependencies across address spaces in the Zope content management system.[18]
- **A**nalyzed specific anomalies in the METS specification (Dubin 2005).
- Reviewed an earlier proposal for repository-like services in networked file systems, considering it in the context of workflow.
- Refocused our attention on the key problem of identifying digital preservation targets: what exactly was to be preserved.

We completed Phase 1 having modified our BECHAMEL Markup Semantics workbench for input and output of object, property, and relation knowledge in serialized RDF, and a proof of concept demonstration for transmission of that knowledge from and to a remote database of RDF triples (Dubin et al, 2009).

4.2 Summary of Semantic Preservation Research during ECHO DEP 2

Progress made during Phase 1 of ECHO-DEP served to crystallize more fundamental problems of resource preservation and description, which were articulated and addressed during Phase 2:

1. Conventional use of the RDF reification vocabulary is based on an understanding that triples stand in a type/instance relationship with "tokens" appearing in RDF documents. But this convention, intended to support provenance documentation, presents puzzles for understanding how a serialized expression can stand in direct relationships with resources referred to by an abstract triple (Dubin, 2007).
2. The integration of RDF-based languages with logic programming tools is guided and constrained by issues of decidability and the tractability of computations. Users of these technologies are invited to use less expressive

[18] http://www.zope.org/

representations, and thereby work within those constraints. Such compromises seem reasonable when considering the roles automated reasoning agents are expected to play by the semantic web community (Dubin and Birnbaum, 2008). But these assumptions are not always appropriate to the particular challenges offered by digital preservation.

3. Formal accounts of digital objects typically characterize them as abstract universals, implying that these objects do not change And yet our discourse about digital objects seems, at least if taken literally, to imply that those objects routinely undergo real change (Renear, Dubin, and Wickett, 2008), and that preservation failures are examples of change.
4. The particular classes of digital object presumed to be the target of digital preservation (e.g., works, texts, editions) are not, by some accounts, types of entities in the strictest sense, but rather roles (Renear and Dubin, 2007). From this perspective, digital preservation is not merely a matter of an entity's continued maintenance of its essential properties over time, but also involves the maintenance of certain crucial relational properties.
5. Both digital preservation targets and certain state-like entities tracked in formal models of provenance (Moreau, et al 2008) are essentially anchored to non-repeating events in time, even though they are not themselves bounded in time and space the way that concrete particulars are understood to be. These quasi-abstract objects would therefore seem to belong in the same realm as social objects (Searle, 1999) or so-called "indicated structures" (Levinson, 1990).

4.2.1 The nature of digital objects: precisely what are we preserving?

For many years there have been two candidates for what, precisely, digital objects are: fully abstract universals (e.g., symbol sequences, graphs, trees, relations, automata) or particular concrete arrangements of matter and energy (on magnetic tape, in fiber optic cable, etc). However, conventional notions of digital object identity, location, and provenance make either type of account problematic. A digital object cannot be identified exclusively with any one of the patterned matter/energy bundles that embodies it. But unlike abstract universals, digital objects are anchored to creation and modification events in time. Some of the properties most salient to preservation (such as an object's having been created by a particular person or modified using a particular computer program) cannot be understood in isolation from these key events.

Indeed, the question of whether digital objects are any type of thing at all depends on key relationships between abstract universals and concrete events. Consider, for example, a binary file's encoding of a digital image. Strictly speaking, the property of encoding some particular image is not necessary to that sequence of bits, but contingent on interpretations that guide the execution of computer software. Categories such as "TIFF," "JPEG," or "digital image file," would therefore seem to be roles played by fully abstract sequences, rather than types in their own right (Guarino and Welty, 2000). If we are to understand digital object classes such as "text file," "Windows executable," and "Java bytecode" as types of things, then the fundamental question is one of what would have the property of being a text file (Windows executable, Java bytecode, etc.) necessarily and not merely contingently. But the troubling answer to that question may be that there is no such thing.

The philosophical literature has, over the past thirty years, featured some intriguing proposals for and discussions about classes of quasi-abstract objects having some properties in common with fully abstract universals and resembling concrete particulars in other respects. These include social objects such as debts and property titles (Searle, 1999; Smith and Searle, 2003) and "indicated structures" for understanding the nature of musical works (Levinson, 1980). But we lack a formal account of such entities that could form the basis of information modeling in digital preservation.

We explored a number of frameworks for formalizing digital preservation targets, including extensive-form games (Dresher 1981) and situation theory (Devlin 1995). However, we were unable to produce any account that squared with the intuitions of theorists like Cheney, Lagoze, and Botticelli (2001) that digital objects undergo changes of state over time. The only plausible alternative seemed to a reductionist strategy that identifies digital resources with abstract universals.

4.2.2 The nature of preservation: what role should models and tools play?

If a preservation model can't reconcile every intuition we have about information resources, it should, at least, serve in the following capacities:

Explanatory

A preservation model should provide a framework in which familiar digital preservation risks can be situated, understood, and classified.

Guiding inference

In earlier writings we present digital preservation as an inference problem, specifically the explication of preservation targets and their properties from evidence in conventional resource descriptions (Dubin, Futrelle, and Plutchak, 2006; Dubin et al, 2009). The required inferences are typically ones that human minds make without conscious effort, but which aren't explicit enough to guide the execution of automated procedures, such as format migrations over large resource collections. A digital preservation model should be precise and formal enough to support automatic or computer-aided inferences.

Informing descriptive practice

Although automatic inference may help software to fill in those semantic gaps that humans bridge without difficulty, a better solution may be to reform descriptive practice to make complicated inference less necessary. A formal preservation model should suggest ways to improve conventional preservation metadata, calling attention to facts that can head off the need for sophisticated deduction if documented directly.

The following section presents a preliminary version of a reductionist model intended to serve in each of these three areas. It is not a complete formalization, still relying to a great extent on word meanings. Most, though not all, of the axioms can be expressed in a description logic that would support limited inferencing with familiar Semantic Web reasoning tools. The section concludes with four scenarios that situate preservation threats in the framework of the model, and point to communication gaps that might be bridged through reformed descriptive or documentation protocols.

4.2.3 Preservation Model, version 1.0

Many, but not all of the axioms below can be expressed in the description logic ALC. Some among those that can govern inferences that we believe are important in preservation, but that can't be carried out by common DL reasoning systems. We therefore express the entire model in first order logic and set-theoretic notation.

We take certain concrete particulars, abstract universals, concrete non-repeating events, and agents as primitive concepts. Concrete particulars include quantities of matter and energy:

- $\forall x(\mathrm{QuantityOfMatter}(x) \rightarrow \mathrm{ConcreteThing}(x))$
- $\forall x(\mathrm{QuantityOfEnergy}(x) \rightarrow \mathrm{ConcreteThing}(x))$

The abstract universals that concern us include all arrangements of symbols (in sequences, graphs, geometric shapes, etc.), patterns of energy/matter, and functions:

- $\forall x(\mathrm{SymbolStructure}(x) \rightarrow \mathrm{AbstractThing}(x))$
- $\forall x(\mathrm{PartialFunction}(x) \rightarrow \mathrm{AbstractThing}(x))$
- $\forall x(\mathrm{PhysicalPattern}(x) \rightarrow \mathrm{AbstractThing}(x))$

Key event types in the model include indications, the selection or determination of an abstract symbol pattern by an agent, and inscriptions, the fixing of a discrete symbol pattern in a tangible medium of expression via positive and reliable techniques (Haugeland, 1985). The model specifies no specific theory of agency, except to specify that certain events have agents:

- $\forall x\forall y(\mathrm{agentOf}(x,y) \rightarrow \mathrm{Event}(x) \land \mathrm{Agent}(y))$
- $\forall x(\mathrm{Indication}(x) \rightarrow \mathrm{Event}(x))$

- $\forall x(\text{Inscription}(x) \rightarrow \text{Event}(x))$

All indication events have at least one agent, and yield one or more indicated symbol structures as objects:

- $\forall x(\text{Indication}(x) \rightarrow \exists y \exists z(\text{agentOf}(x,y) \wedge \text{objectOf}(x,z)))$
- $\forall x \forall y(\text{objectOf}(x,y) \rightarrow \text{Event}(x) \wedge \text{SymbolStructure}(y))$

Some indication events and all inscription events employ a mapping, which is a partial function from the set of all abstract symbol structures to either itself (indications) or from/to the set of abstract physical patterns (inscriptions). An example of the former would be a function from bit sequences to EBCDIC character strings, and an example of the latter would be a function from EBCDIC characters to hole punch patterns in cardboard Hollerith cards:

- $\forall x(\text{SymbolMapping}(x) \rightarrow \text{PartialFunction}(x))$
- $\forall x(\text{PatternMapping}(x) \rightarrow \text{PartialFunction}(x))$
- $S \equiv \{x: \text{SymbolStructure}(x)\}$
- $P \equiv \{x: \text{PhysicalPattern}(x)\}$
- $\boldsymbol{I}(\text{SymbolMapping}) \equiv \{f: S' \rightarrow S, S' \subseteq S\}$
- $\boldsymbol{I}(\text{PatternMapping}) \equiv \{f: S' \rightarrow P, S' \subseteq S\} \cup \{f: P' \rightarrow S, P' \subseteq P\}$
- $\forall x \forall y(\text{mappingOf}(x,y) \rightarrow \text{Event}(x) \wedge (\text{SymbolMapping}(y) \vee \text{PatternMapping}(y)))$

Some mappings are known and available to agents of preservation transactions, while others are unknown. For example, the standard mapping from UTF-8 encoded octet sequences to UCS character sequences is known, but mappings that can correctly govern interpretations of the Voynich Manuscript and the Phaistos Disc are unknown, may never be known, and might not exist:

- $\forall x(\text{KnownMapping}(x) \rightarrow \text{SymbolMapping}(x) \vee \text{PatternMapping}(x))$

Transliterations are a subclass of indication, in which one symbol structure (the source) is mapped to another (the object) via a symbol mapping. The source structure is therefore understood as a subclass of basis, and transliteration is one way that a symbol structure can be derived from another. The source, mapping, and object are assumed to be unique for any particular transliteration event:

- $\forall x(\text{Transliteration}(x) \rightarrow \text{Indication}(x))$
- $\forall x(\text{Transliteration}(x) \rightarrow \exists y \exists z \exists w(\text{SymbolMapping}(y) \wedge \text{sourceFor}(x,z) \wedge \text{mappingOf}(x,y) \wedge \text{objectOf}(x,w) \wedge \langle z,w \rangle \in y))$
- $\forall x \forall y(\text{basisFor}(x,y) \rightarrow \text{SymbolStructure}(y) \wedge \text{Indication}(x))$
- $\forall x \forall y(\text{sourceFor}(x,y) \rightarrow \text{basisFor}(x,y) \wedge \text{Transliteration}(x))$
- $\forall x \forall y \forall z((\text{basisFor}(x,y) \wedge \text{objectOf}(x,z)) \rightarrow \text{derivedFrom}(y,z)$
- $\forall x \forall y \forall z((\text{Transliteration}(x) \wedge \text{sourceFor}(x,y) \wedge \text{sourceFor}(x,z)) \rightarrow y=z)$
- $\forall x \forall y \forall z((\text{Transliteration}(x) \wedge \text{objectOf}(x,y) \wedge \text{objectOf}(x,z)) \rightarrow y=z)$
- $\forall x \forall y \forall z((\text{Transliteration}(x) \wedge \text{mappingOf}(x,y) \wedge \text{mappingOf}(x,z)) \rightarrow y=z)$

By a *Digital Resource* we mean an abstract symbol structure that has been the object of some concrete indication event. This might be a transliteration event, or some other type of indication, such as an act of authorship or an adaptation (e.g., a translation):

- $\forall x(\mathrm{DigitalResource}(x) \leftrightarrow (\mathrm{SymbolStructure}(x) \land \exists y(\mathrm{Indication}(y) \land \mathrm{objectOf}(x,y))))$

An inscription event fixes a symbol pattern to some particular quantity of matter or energy (the medium). An inscription expresses only one symbol structure directly, though others may be encoded (and therefore preserved) indirectly via preceding transliteration events:

- $\forall x(\mathrm{Inscription}(x) \rightarrow \exists y \exists z \exists w \exists v \exists u(\mathrm{agentOf}(x,y) \land \mathrm{arrangementOf}(x,z) \land \mathrm{mediumOf}(x,w))) \land \mathrm{objectOf}(x,v) \land \mathrm{mappingOf}(x,u) \land \langle v,z \rangle \in u$
- $\forall x \forall y(\mathrm{mediumOf}(x,y) \rightarrow \mathrm{Inscription}(x) \land (\mathrm{QuantityOfMatter}(y) \lor \mathrm{QuantityOfEnergy}(y)))$
- $\forall x \forall y(\mathrm{arrangementOf}(x,y) \rightarrow \mathrm{Inscription}(x) \land \mathrm{PhysicalPattern}(y))$
- $\forall x \forall y \forall z((\mathrm{Inscription}(x) \land \mathrm{mappingOf}(x,y) \land \mathrm{mappingOf}(x,z)) \rightarrow y=z)$
- $\forall x \forall y \forall z((\mathrm{Inscription}(x) \land \mathrm{objectOf}(x,y) \land \mathrm{objectOf}(x,z)) \rightarrow y=z)$
- $\forall x \forall y \forall z((\mathrm{Inscription}(x) \land \mathrm{arrangementOf}(x,y) \land \mathrm{arrangementOf}(x,z)) \rightarrow y=z)$
- $\forall x \forall y \forall z((\mathrm{Inscription}(x) \land \mathrm{mediumOf}(x,y) \land \mathrm{mediumOf}(x,z)) \rightarrow y=z)$

A conformance relation is understood to hold between a quantity of matter and a physical pattern into which it is arranged. The conformance relation can cease to obtain if the matter or energy is rearranged into some other pattern.

- $\forall x \forall y(\mathrm{conforms}(x,y) \rightarrow (\mathrm{PhysicalPattern}(y) \land (\mathrm{QuantityOfMatter}(x) \lor \mathrm{QuantityOfEnergy}(x))))$

By an *Epigraph* we mean a quantity of matter or energy that has been the medium of some inscription event and continues to conform to the inscribed arrangement:

- $\forall x(\mathrm{Epigraph}(x) \leftrightarrow ((\mathrm{QuantityOfMatter}(x) \lor \mathrm{QuantityOfEnergy}(x)) \land \exists y \exists z(\mathrm{Inscription}(y) \land \mathrm{mediumOf}(y,x) \land \mathrm{arrangementOf}(y,z) \land \mathrm{conforms}(x,z)$

An epigraph directly *preserves* a symbol structure that was the object of its inscription, provided that the inscription's pattern mapping has a known inverse that will allow an agent to recover the object:

- $\forall w \forall x \forall y \forall z[(\mathrm{Epigraph}(x) \land \mathrm{mediumOf}(y,x) \land \mathrm{arrangementOf}(y,z) \land \mathrm{objectOf}(y,w)) \rightarrow ((\exists v(\mathrm{KnownMapping}(v) \land \langle z,w \rangle \in v) \rightarrow \mathrm{preserves}(x,w))]$

If an epigraph preserves the object of a transliteration event and a known mapping enables the recovery of the transliteration source, then the epigraph indirectly preserves the source:

- $\forall w \forall x \forall y \forall z[(\mathrm{preserves}(x,w) \land \mathrm{transliteration}(y) \land \mathrm{sourceFor}(y,z) \land \mathrm{objectOf}(y,w) \rightarrow (\exists v(\mathrm{KnownMapping}(v) \land \langle w,z \rangle \in v) \rightarrow \mathrm{preserves}(x,z)))]$

This model identifies digital resources with fully abstract symbol structures, and offers no basis for identity conditions apart from those the abstract objects themselves supply. As abstract objects, resources never undergo any change of state. Digital preservation risks that seem to presuppose changes of state or resources identifiable at levels of abstraction not included in the model therefore need to be reinterpreted as procedural, documentation, or communications failures. The following examples may serve to illustrate:

1. Suppose that the four character string "LOL!" is sent in two different SMS text messages one hour apart by two different people. In the context of our preservation model we have two different indication events, one for each sender. But the abstract string selected is exactly the same in both cases: our digital resource is an uncreated and immutable abstract string that has the status of a digital resource for at least two different reasons. Any information serving to distinguish these two messages from each other (in, e.g., an OA or DA field of the SMS PDU) would have to be understood either as metadata attached to the resource or else part of a distinct derived resource (the entire PDU as a bit string) selected in a separate indication event. In the latter case, the seven-bit encoding of "LOL!" in the UD field of the PDU would have been transliterated from a prior expression of the message in a cell phone memory buffer.
2. Consider a scenario in which the only media on which a digital resource is recorded is damaged beyond any hope of recovering the data. Although the quantity of matter serving as the storage medium continues to exist after the damage, it fails to conform to the relevant physical arrangement, and therefore ceases to be an epigraph of the digital data. If there exists no epigraph directly or indirectly preserving that data, then the digital resource is not preserved.
3. Consider a transliteration of a string of Cyrillic characters from UTF-16 encoded UCS into eight bit ISO 8859-5. According to our model, any epigraph preserving the latter string would also preserve the original Cyrillic text, as long as mappings from relevant physical patterns to bits and from 8859-5 octets back to Cyrillic characters continue to be known. However, that's not to say the recovery of the original resource will be easy (or even possible) absent metadata documenting the encoding of the text: some probabilistic analysis of the file contents might be necessary to discover the correct encoding.
4. Suppose that a file expressed in a legacy database format begins with the bit string 0100100101001001, and is for that reason misidentified as a TIFF image. Suppose that a crude (and lossy) TIFF to JPEG conversion program is naively applied, resulting in a stream that is not a valid JPEG and that the original database file is erased. Call the original database stream a and the stream emerging from the conversion b. Trivially, there must exist a partial function consisting only of the ordered pair <b,a>, and since many abstract

> symbol systems are productive, an infinite number of other mappings will exist that also contain that pair. But none of those mappings are likely to be known and available to preservation agents. In fact, it's unlikely that any mapping capable of governing a successful recovery of the original database file contents is known. Therefore, however many quantities of matter directly or indirectly preserve the second (non-JPEG) stream, the original database file will not have been preserved.

5. Extracting Metadata for Preservation Project

> *This portion of the ECHO DEPository research grant was completely performed in phase two of the grant, and involved researchers at the University of Illinois, the Online Computer Library Center (OCLC), and the University of Maryland. A paper addressing this work was presented at the 2009 Chicago Colloquium on Digital Humanities and Computer Science (Godby, Hswe, Jackson, Klavans, Ratinov, Roth, & Cho, 2009).*

In the past twenty years, the problem space of automatically recognizing, extracting, classifying, and disambiguating named entities (e.g., the names of people, places, and organizations) from digitized text has received considerable attention in research produced by the library, computer science, and computational linguistics communities. However, linking the output of these advances with the library community continues to be a challenge. In this work, we addressed developed, evaluated and linked Named Entity Recognition (NER) and Entity Resolution with tools used for search and access. Name identification and extraction tools, particularly when integrated with a resolution into an authority file (e.g., WorldCat Identities, Wikipedia, etc.), can enhance reliable subject access for a document collection, improving document discoverability by end-users (Cucerzan, 2007).

In the context of historical documents, the ability to find out who knew whom and why they were associated, in addition to whether the individuals are actually the ones the user is seeking, cultivates a potential for further, value-adding analysis of the documents' content. Discerning who's who in a digital resource collection is increasingly of interest to archivists, curators, and humanities scholars. The Perseus Digital Library[19] has Named Entity Search Tools that mine its collections for people, places, and even dates. The Metadata Offer New Knowledge (MONK) project[20] offers a workbench for textual analysis on multiple levels, including a tool for recognizing and extracting named entities in its collections (which consist of works of eighteenth- and nineteenth-century American literature and works by William Shakespeare). Named-entity extractors can also be found in cataloging utilities,

[19] http://www.perseus.tufts.edu/hopper/

[20] http://www.monkproject.org/

such as the Computational Linguistics for Metadata Building (CLiMB) Toolkit[21] , which addresses the "subject metadata gap" in visual resources cataloging by increasing subject access points for images of art objects (Klavans, Abels, Lin, Passonneau, Sheffield, & Soergel, 2009).

The problem of name disambiguation and identity resolution is made especially acute when many entities share the same name. Suppose a historian is seeking new insights about the assassination of John Kennedy. A Google search reveals that there are more than a few men named John Kennedy; the surname Kennedy itself is popular. The texts excerpted in Table 1 describe about various Kennedys. To identify the relevant resources, the scholar would have to sift through search results one by one, a tedious task calling for automation. What would it take?

[21] http://www.umiacs.umd.edu/~climb/

Document 1: “Composer and conductor John Kennedy is a dynamic and energetic figure in American music. Recognized for his artistic leadership, imaginative programming, audience development, and expertise in the music of our time, Kennedy has conducted celebrated performances of opera, ballet, standard orchestral and new music. His own compositions, from operas to chamber works, are praised for their new lyricism and luminous sound.”[22]
Document 2: “In 1953, Massachusetts Sen. John F. Kennedy married Jacqueline Lee Bouvier in Newport, R.I. In 1960, Democratic presidential candidate John F. Kennedy confronted the issue of his Roman Catholic faith by telling a Protestant group in Houston, “I do not speak for my church on public matters, and the church does not speak for me.’”[23]
Document 3: “John Kennedy was elected without opposition to his third term as State Treasurer in 2007. As Treasurer, he manages the state’s $5 billion bank account including the investment of $3 billion in trust funds. He also oversees local and state bond issues and returns millions of dollars in unclaimed property each year. Prior to his position as Treasurer, Mr. Kennedy served as Secretary of the Department of Revenue, Special Counsel to Governor Roemer and Secretary of Governor Roemer’s Cabinet.”[24]

Table 1. Three texts about men named Kennedy.

The ideal software process would have to perform three tasks well enough to satisfy a discerning human judge. First, it would have to recognize the names. All name recognition software works by ingesting a string of text, such as the first sentence in the second document, and separating the names (Massachusetts, Sen. John F. Kennedy, Jacqueline Lee Bouvier, Newport, and R.I.) from the non-names (In, 1953, and married). This is a non-trivial task because the recognizer has to be smart enough to pick out names consisting of text strings that span more than one word, such as Jacqueline Lee Bouvier. It must also skip over the periods that indicate abbreviations (as in Sen. John. F. Kennedy or R.I.), but not those at the end of a

[22] Quoted from the website about John Kennedy: http://www.johnkennedymusic.com/about.html. Retrieved December 15, 2009.

[23] Quoted from the Wikipedia entry on John F. Kennedy: http://en.wikipedia.org/wiki/John_F._Kennedy. Retrieved December 15, 2009.

[24] Quoted from the website for John Neely Kennedy: http://www.treasury.state.la.us/Home%20Pages/TreasurerKennedy.aspx. Retrieved December 15, 2009.

sentence. Second, the recognizer must categorize the names. In the sample texts, all of the name strings containing the word Kennedy refer to people, although this will not always be true because the system will eventually encounter a text containing the organization name such as John F. Kennedy School of Government or a place name such as Kennedy Airport. It could also encounter strings that in some context are names, and in others are not, such as the first word in the sentence "Begin was the prime minister of Israel." Categorization effectiveness is a function of the diversity and extent of the training data supplied and of the algorithmic approach used. Finally, the software procedure must perform the most difficult task of all: assigning the real-world referents to the name strings. To help the scholar, the software would have to distinguish the John Kennedy from everyone and everything else named Kennedy, a task known as name disambiguation or identity resolution.

Because our project team has many librarians, we are interested in supporting research and scholarship like that of the hypothetical historian. An automated name recognizer paired with an identity resolver would support this goal and many others, including those that are central to the mission of libraries. For example, the output from these programs could be used to create more responsive interfaces for the discovery and retrieval of library materials. Or it could supply input to improved versions of resources that authoritatively describe the places, the people, and their inventions discussed in the published record, as well as the authors themselves.

Since there is no question that name recognition and identity resolution software would be key technologies for many applications enlisted in the service of preserving cultural memory, it is more interesting to ask why they haven't been pressed into service. The usual answer is that these programs, although incorporated to some degree in multiple commercial products, are not ready for full-scale deployment. They may not be freely available or are difficult to use out of the box; processing time is too slow; the output has too many errors; and only name recognition, not entity resolution, is mature enough for serious consideration. But it's also undeniable that the output from these tools is already good enough for some library applications. To unleash their potential, researchers in the library community need to match this new technology with use cases that tolerate the current state of the art; form partnerships with the computer-science researchers engaged in front-line research in name recognition and identity resolution; and define realistic goals for future development.

To address these issues, we proposed the Extracting Metadata for Preservation (EMP) Project, funded by the National Digital Information Infrastructure and Preservation (NDIIPP) Program. As a collaboration among the University of Illinois at Urbana-Champaign, OCLC, and the University of Maryland, EMP researchers bring multidisciplinary perspectives from the library, computer science, and linguistics communities to the problem of high-quality identification and disambiguation of names. Our work has three goals: 1) to advance the state of the art in automated name identification and disambiguation; 2) to link the outputs of these programs to longstanding efforts in the library community to manage names and identities in the published record; and 3) to lower the barrier of access to these tools.

5.1 Related Research

Named Entity Recognition (NER) has been a key subject for researchers interested in accurate content extraction, information extraction, and information retrieval. Due to the centrality of personal names, places, dates, organizations and other named entities (NEs) in characterizing the topics in a document, audio or video clip, the quest for exactness in tokenizing these items has a long history. One of the earliest efforts to measure occurred at the Message Understanding Conferences (MUC), a series of workshops funded by the Defense Advanced Research Projects Association (DARPA). Projects funded by MUC participated in what are fondly called "computational linguistic bake-off's", where each system was run over a set of common data with results being submitted for evaluation by an independent set of evaluators through technology developed at the National Institute of Standards and Technology (NIST). The Named Entity task for MUC-6, held in 1995, consisted of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are "unique identifiers" of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages). This task was intended to be of direct practical value (in annotating text so that it can be searched for names, places, dates, etc.) and an essential component of many language-processing tasks, such as information extraction (Grishman & Sundheim, 1995).

More recent approaches use a variety of techniques. In 2003 an overview of methods was provided at a workshop conducted by the annual Conference on Natural Language Learning (CoNLL)[25], supported by the Special Interest Group on Natural Language Learning of the Association for Computational Linguistics (Sang & DeMeulder, 2003). This reflects the current belief in the natural language processing and information extraction communities, that machine learning techniques, rather than programmed (rule-based) systems, are necessary in order to address the NER problem (and many other related problems) (Klavans & Resnik, 1996). Despite the emphasis on statistical machine learning techniques, most of the participants have attempted to use information other than the available training data, such as gazetteers and un-annotated data.

The most frequently applied techniques in the CoNLL-2003 shared task were sequential classifiers of different sorts. At that time, one of the most popular sequential classifiers was the Maximum Entropy Model (MEM), but several other sequential classifiers, such as Hidden Markov Models and Conditional Markov

[25] More information is available at: http://www.cnts.ua.ac.be/conll2003/ner/. Also referred to in this paper as the "CoNLL tagging scheme."

Models (Finkel, Grenager, & Manning, 2005) also were used. Many other machine learning approaches—including connectionist approaches, robust risk minimization, transformation-based learning, and support vector machines—were used for this problem, but it is clear today that architectural issues and features are the most important decisions, more than the specific training algorithm used.

One of the most complex tasks within the NER area is that of identifying nested entities. For example, "Columbia University in the City of New York" is an organization; however, the nested entity "City of New York" is a location, as is the entity nested within the nest, "New York." Many corpus designers have chosen to avoid the issue of nesting entirely and have annotated only the topmost entities. CoNLL (Sang & DeMeulder, 2003), MUC-6, and MUC-7 NER corpora, composed of American and British newswire, are all flatly annotated. A partial reason for this is that the NER task arose in the context of the MUC workshops, as small chunks of text which could be identified by finite state models or gazetteers. This then led to the widespread use of sequence models—first hidden Markov models, then conditional Markov models (Borthwick, 1999), and, more recently, linear chain conditional random fields (CRFs) (Lafferty, McCallum, & Pereira, 2001). None of these are able to model nested entities. Moreover, in essentially all sequential models it is often computationally difficult to represent non-local dependencies, which are often important in NER. This is one reason the approach used in this research (Ratinov & Roth, 2009) is not based on sequential classifiers but, rather, on state-of-the-art classifiers, which allows us to flexibly include non-local information.

5.2 The Name Extractor Tool

The EMP project uses a Named Entity Tagger[26], developed at the Cognitive Computation Group at UIUC (Ratinov & Roth, 2009). This NER, based on Roth's research group's earlier machine learning modeling language, Learning Based Java (LBJ)[27], was shown to be the best performing tool available today and its efficiency allows it to be used as part of applications that process large amounts of data. It extracts and labels non-nested named entities into four categories: locations (LOC), persons (PER), organizations (ORG), and miscellaneous names of human-created artifacts (MISC).

The algorithm incorporates a general model that learns from examples to identify named entities and classify them. It works in two stages. The baseline model makes a first cut by classifying the input text greedily left to right, using features that include, but are not limited to, the previous two tokens, the previous two classifications, and capitalization features. Most notably, the system does not use Part-Of-Speech tagging or shallow parsing information, which are common in other

[26] An online demo of this software is available at http://l2r.cs.uiuc.edu/~cogcomp/LbjNer.php.

[27] http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=LBJ

NER taggers. The second stage makes use of nonlocal features and features that exploit external knowledge. The classification model underlying the LBJ Named Entity Tagger is a regularized averaged perceptron[28] algorithm.

The two additional feature types added to the LBJ NER, along with other design decisions, account for its performance, which exceeds that of other state of the art tools and provides a necessary ability to adapt well to text from multiple domains and genre. Both feature types rely on automatically constructed evidence collected as part of the learning process. First, the system uses nonlocal features, such as the ratio of Named Entity types assigned to the current token previously in the text and context aggregation. By doing so, it makes use of the two-stage predication, where the first model is used to classify the text, while another model, similar in nature to the first, corrects the predictions to make them consistent within a document. Second, the system uses word class models and massive gazetteers automatically extracted from the online resource Wikipedia.[29]

Consider, for example, the text in Table 2. The system may incorrectly classify the first instance of "Blinker" at the first level of inference, but it will correct the prediction at the second level of inference by seeing that "Blinker" was a part of the expression "Reggie Blinker", labeled as person ("PER"). Furthermore, the system will use the knowledge extracted from Wikipedia, which states that "Udinese", "Sheffield Wednesday", "Liverpool", and "Feyenoord" are football (soccer) clubs -- a kind of organization ("ORG"). The system will correctly label the second instance of "Wednesday," since the expression "Sheffield Wednesday" was labeled as an ORG previously in the text. It is also important to note that the system uses the algorithm with large amounts of unlabeled text to abstract away words to a word class model, thus avoiding problems of data sparseness common in Natural Language Processing (NLP). For example, given the sentence containing “FIFA slapped,” the system knows that slapped is used in similar contexts as “devised, re-imposed, manifested, commissioned, authorized, imposed, etc,” helping the system to label "FIFA" as an ORG.[30]

[28] A perceptron is an “On-line, mistake driven, additive update rule. A perceptron updates the weights in a target node by adding to them a learning rate that is a function of the type of mistake made (either positive or negative) and the strengths of features in the example” (Carlson, Cumby, Rizzolo, Rosen, & Roth, 2004).

[29] This process is called 'wikification,' and is demonstrated by the University of North Texas facility at http://wikifyer.com/.

[30] More details of the tool's operational principles may be found in Ratinov & Roth 2009).

SOCCER - [PER BLINKER] BAN LIFTED .
[LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make his [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday. [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord].

Table 2. Text displaying the annotated output of the LBJ Named Entity Tagger.

In addition to the extensive evaluation described in the CoNLL 2009 presentation by Ratinov and Roth, we also assessed how well the LBJ Named Entity Tagger performs in comparison with other state-of-the-art name extractor applications used in the library community. Besides the LBJ tagger employed in our project, two other tools were assessed: ClearForest Gnosis (ClearForest)[31], which is a FireFox add-on application that semantically processes webpages, linking named entities to further information about them; and the Stanford Named Entity Recognizer (NER), developed by the Stanford Natural Language Processing Group using a Character-based Maximum Entropy Markov Mode (MEMM), which is also implemented in Java.[32] For the additional evaluation, we selected five text samples taken from diverse domains, ran the samples through each tool, and compared the raw performance of the results. We also engaged a human annotator to tag named entities in each text sample and compared the human-generated results with those obtained from evaluation of the aforementioned three NER tools. It is important to note that in all cases addressed here the tool was evaluated on text taken from domains that are vastly different from the domain it was trained on. In principle, when one wants to use such a tool in a different domain, the best course of action is to re-train the tool on the target domain. The results here, therefore, should also be taken as evidence of the robustness and adaptability of the tool.

For name occurrences ("mentions") that were exactly matched, the F-scores for the LBJ tagger on the five text samples ranged from 47.83% to 78.99%, depending on the domain; for partially matched mentions, the F-scores ranged from 60.13% to 85.71%. The closest competitor, ClearForest, had F-scores for exactly matched mentions that ranged from 36.14% to 61.73%; for partially matched mentions, the

[31] https://addons.mozilla.org/en-US/firefox/addon/3999

[32] Available at: http://nlp.stanford.edu/software/CRF-NER.shtml. The Stanford NER is also evaluated in (Ratinov & Roth, 2009).

F-scores for ClearForest ranged from 42.77% to 75.86%. In the version evaluated, the LBJ NER tool was tuned to yield the best F1 score, which is the harmonic average of recall and precision, although it is possible to tune it to emphasize one over the other. In general, a high precision rate is often important in dealing with extremely large collections, since the latter would be likely to yield more errors, and thereby waste the user's time. High recall rates, though, reflect the coverage of the tool—the percentage of entities identified—and are desirable where the search must be exhaustive, such as in research or legal applications. In general, with the version evaluated ClearForest had slightly higher precision but significantly lower recall (that is, it identified significantly fewer entities). One lesson from this evaluation that we intend to act on is to simplify the ability of a user to retrain the LBJ NER tool on a target domain, and to allow a user to easily trade recall and precision.

5.3 Resolving Identities

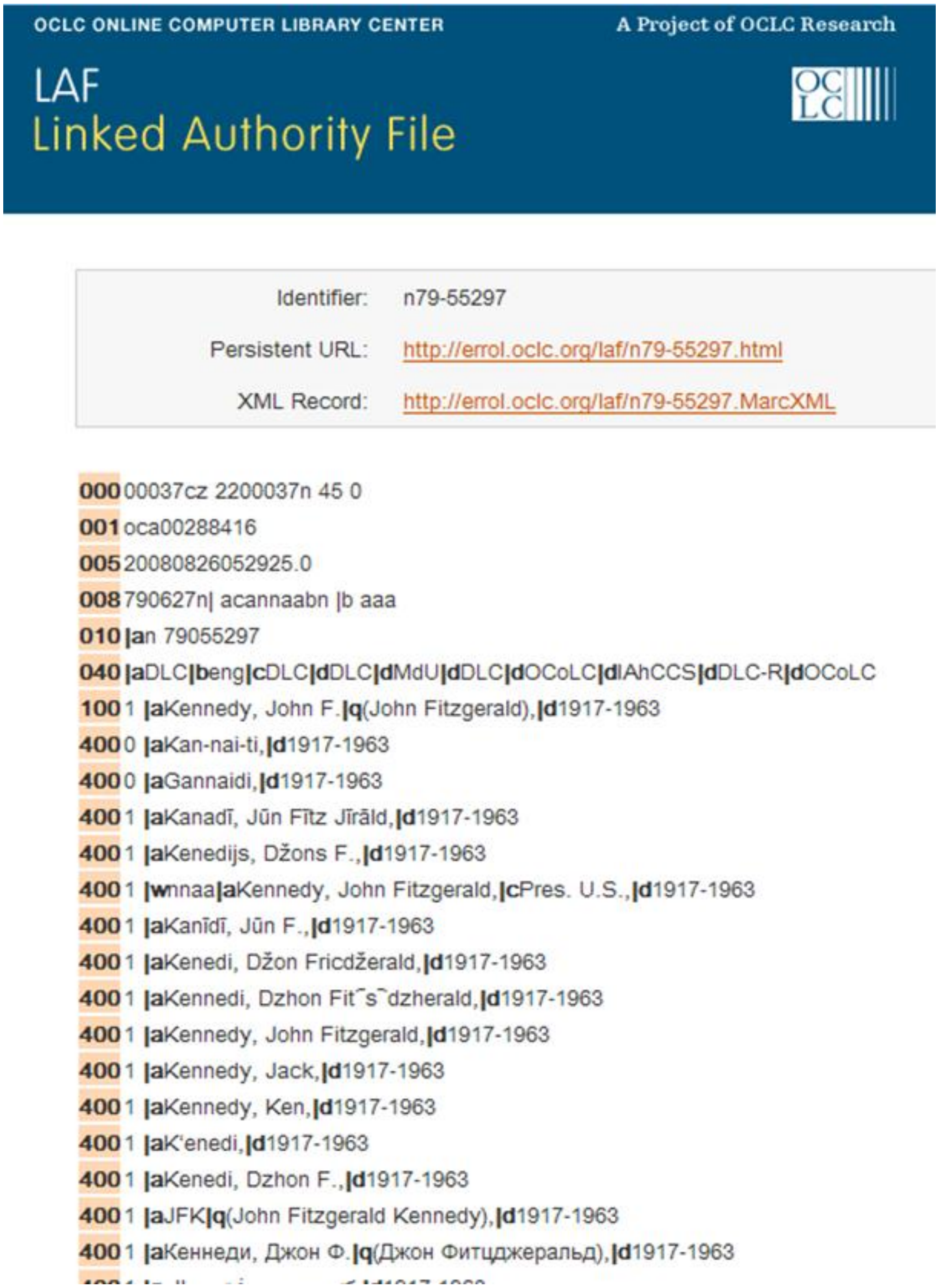
As we said above, entity recognition is only the first part of the problem of capitalizing on the rich information associated with names in unstructured text. The second is identity resolution: determining which person, place, or concept in the real world the extracted name refers to. This is a classic problem in the philosophy of language (Kripke, 2000). In a nutshell, identity resolution requires the help of an authority who can step outside the text and link the name with the appropriate referent—such as a mother who names her child "John Fitzgerald Kennedy", a public official who witnesses this act, or a journalist who writes about it. This link then needs to be fixed so that it remains constant over time, persisting even into eras when the named entity has passed out of living memory. Thus, if the name-referent link is robust, 23rd-century readers of a book published in 1966 about the assassination of John Fitzgerald Kennedy will understand that the book is about the American president who was elected in 1960, just as their counterparts in the 20th century did.

Since the creation of a name-referent link is a vexing problem for philosophers and is occasionally challenging for human readers, it would appear intractable for a software algorithm that does not have access to the world beyond a set of input texts. Except for the people, places, and things encountered in their everyday experience, humans don’t have this access, either. But they still manage to understand texts like those excerpted in Table 1. We can infer that since relatively few people are personally acquainted with the composer, the 35th American president, or the state treasurer of Louisiana (the examples presented in Table 1 above), they grasp the meaning of these texts by consulting identity resolution authorities—textbooks or other works of nonfiction, documentary films, encyclopedias, or their own memories of these works—who describe the identity behind the name in enough detail to establish a proxy reference.

Algorithms that attempt to resolve identities also consult a resolution authority to establish the identities of the various people named Kennedy in texts such as the ones we have described. Stated more formally, the problem to be solved has three parts. First, name occurrences are extracted from the text, such as John Kennedy, or simply Kennedy. Second, a software process must match the name occurrences against those found in an identity resolution authority. This task is easy if the name occurrence is unusual and has only one entry in the authority. But more typically, the name is ambiguous and has multiple representations, which makes a third step necessary: generating candidates from the identity resolution authority and selecting the correct one, a task that usually requires that the input text be mined for clues about the identity of the name occurrence, such as birth and death dates for personal names, or city and country names for places.

So what is a good identity resolution authority for a software process? Computer scientists argue that Wikipedia is appealing because it is a high-quality edited text that is freely available. It has a relatively large coverage (over two million entities as of August 2009) and is frequently updated by human annotators who enhance the hyperlink structure. In particular, the most important named entities mentioned in Wikipedia articles are linked to the corresponding Wikipedia pages, which are also annotated with a list of human-created categories. These features allow us to obtain statistics, such as how often a given set of tokens refers to a given Wikipedia page; how often two Wikipedia concepts appear in the same Wikipedia page; and how the texts are associated with abstract Wikipedia categories. These statistics permit the construction of expressive disambiguation models. Ratinov and Roth are developing a disambiguation system ("wikifier") that assigns the correct Wikipedia entries to named entities and concepts identified in blogs and texts retrieved by standard information retrieval algorithms. Their system builds on the work of researchers who attempted to enrich the hypertext structure of Wikipedia by expanding the list of named entities that link to the corresponding articles, such as (Cucerzan, 2007) or (Mihalcea & Csomai, 2007).

The librarians on the EMP team have proposed the use of library authority files for identity resolution. Typically created by national libraries to establish unambiguous references to the people, places, and topics represented in the published record, library authority files are highly encoded and designed for machine processing. Figure 7 shows a portion of the record for John Fitzgerald Kennedy from the Library of Congress Name Authority File. The various fields in the record supply birth and death dates, alternative forms of his name, associated subjects, and the coded names of the agencies that vouch for the accuracy of this information.



OCLC ONLINE COMPUTER LIBRARY CENTER

A Project of OCLC Research

LAF
Linked Authority File

Identifier:	n79-55297
Persistent URL:	http://errol.oclc.org/laf/n79-55297.html
XML Record:	http://errol.oclc.org/laf/n79-55297.MarcXML

```
000 00037cz 2200037n 45 0
001 oca00288416
005 20080826052925.0
008 790627n| acannaabn |b aaa
010 |an 79055297
040 |aDLC|beng|cDLC|dDLC|dMdU|dDLC|dOCoLC|dIAhCCS|dDLC-R|dOCoLC
100 1 |aKennedy, John F.|q(John Fitzgerald),|d1917-1963
400 0 |aKan-nai-ti,|d1917-1963
400 0 |aGannaidi,|d1917-1963
400 1 |aKanadī, Jūn Fītz Jīrāld,|d1917-1963
400 1 |aKenedijs, Džons F.,|d1917-1963
400 1 |wnnaa|aKennedy, John Fitzgerald,|cPres. U.S.,|d1917-1963
400 1 |aKanīdī, Jūn F.,|d1917-1963
400 1 |aKenedi, Džon Fricdžerald,|d1917-1963
400 1 |aKennedi, Dzhon Fit˜s˜dzherald,|d1917-1963
400 1 |aKennedy, John Fitzgerald,|d1917-1963
400 1 |aKennedy, Jack,|d1917-1963
400 1 |aKennedy, Ken,|d1917-1963
400 1 |aK‘enedi,|d1917-1963
400 1 |aKenedi, Dzhon F.,|d1917-1963
400 1 |aJFK|q(John Fitzgerald Kennedy),|d1917-1963
400 1 |aКеннеди, Джон Ф.|q(Джон Фитцджеральд),|d1917-1963
```

Figure 7. The Library of Congress authority record for John Fitzgerald Kennedy.[33]

[33] http://errol.oclc.org/laf/n79-55297.html

In the past five years, classification experts in the library community have recognized the need to create authority files that span national and linguistic boundaries. One outcome is the Virtual International Authority File[34], a collaborative effort that merges authority files from thirteen national libraries. Another example is OCLC's WorldCat Identities[35], a Web-accessible collection with 27 million pages about personal names[36], which have been populated with links and other data obtained from multiple authority files, Wikipedia, and collections of bibliographic records—in particular, OCLC's database of 158 million records representing records contributed by 71,000 libraries worldwide. Since these resources are automatically compiled, they must also rely on identity resolution algorithms that extract name occurrences and select the correct identity from a list of candidates. But since the authority file data is highly encoded and the scope is restricted to names represented in the published record, it is relatively easy to discover distinctive information such as the names of works an author has published. In the next section, we discuss an extended example that illustrates the use of authority files for identity resolution.

At present, the EMP project team is debating how to reconcile these two approaches to identity resolution. The team's computer scientists argue that the library authority files contain data that is too sparse for algorithms tuned for the rich unstructured text of Wikipedia. Or that Wikipedia is comprehensive, while the library authority files are restricted to the published record. It is also clear, however, that the two types of resources are complementary. If the goal is to identify the names of authors extracted from text obtained from the open Web, the correct resolution is more likely to come from WorldCat Identities than from Wikipedia, which currently has fewer than 125,000 articles about authors. At the same time, WorldCat Identities can be probably be enhanced by algorithms that work on unstructured text: they promise to locate authors who are well-known and influential yet not represented in the published record, since they speak only through blogs or websites that have "gone viral."

5.3.1 Wikifier extensions to the UIUC NER tool

Wikification is the task of identifying and linking concept mentions (expressions) in text to their referent Wikipedia pages. Wikipedia's rich descriptions and link structure provide important clues that aid in accurately detecting target entities. In Wikification, prior information about which entities a string tends to refer to is very indicative. Wikification systems typically output a single best disambiguation for

[34] http://viaf.org/

[35] http://orlabs.oclc.org/Identities/

[36] WorldCat Identities also has 7 million pages about corporate names and 14,000 subject names (Ralph LeVan, personal communication).

each surface form. A primary challenge in Wikification is the disambiguating of concepts at a fine level of granularity.

We follow the general wikification approach of linking all the named entities, as opposed to linking selected instances of a broader set of expressions, mimicking the link structure of Wikipedia**.** We here refer to the various textual substrings that may refer to an entity as surface forms, the references $\mathbf{s_i}$ that do correspond to some Wikipedia entity as mentions, and Wikipedia pages titles c_i as concepts. We follow Wikipedia notation, where `[[c|s]]` denotes surface form s linked to a concept c. For example, `[[Chicago_(album)|Chicago_II]]` means that the surface form "Chicago_II" is hyperlinked to Wikipedia page http://en.wikipedia.org/wiki/Chicago_(album).

Our Wikification work leverages information mined from Wikipedia to perform disambiguation, specifically;

- the information type for a concept c and surface string s,
- the number of Wikipedia pages which refer to c divided by the total number of Wikipedia pages,
- the number of hyperlinks to c which used s as a surface form divided by the total number of hyperlinks to c,
- the number of hyperlinks with surface form s linked to the concept c as opposed to other concepts,
- the number of pages where s is hyperlinked divided by the total number of pages which contain s,
- the contexts surrounding all hyperlinks to c throughout Wikipedia,
- the text of the article c,
- the articles which c links to,
- the articles which link to c, and
- the categories of the article c.

For each Wikipedia concept, we record the number of pages that contained a link to it. This property, denoted by `P(c)` corresponds to concept prevalence, or the prior probability of a concept to appear. For each concept c we go over all the hyperlinks pointing to c, and record the surface forms which were used to anchor the concept. We use this information to build the conditional distribution `P(s|c)` of using a surface form s to represent concept c. We also build a reverse distribution `P(c|s)` of concepts which a given surface form s can refer to. For each surface form s, `P(link|s)` denotes the number of pages which contained a hyperlink anchored by s as opposed to pages that contain only raw text version of s. `P(link|s)` (i.e., he number of pages which contained a hyperlink anchored by a particular surface concept s, as opposed to pages that contain only raw text version of s) closely correlates to the degree to which s refers to a named entity. We use this feature extensively to recover from erroneous delineations in cases where other NER tools fail.

For each Wikipedia article corresponding to a concept c, we extract the list of concepts which are linked to c and linked from c. We use this information to estimate concept relatedness. For example, if many articles contain hyperlinks both to Chicago (band) and Rock music, we conclude that the two concepts are closely related. For similar purposes, we extract the categories of the concepts. Finally, we extract the text of the concept, and all the text surrounding the hyperlinks to the concept. Both resources are lexical, where the former relies on the text of the article describing the concept, and the latter aggregates the local context within which the concept was mentioned in other articles.

In testing the wikifier, we used the EMP NER tagger to identify the named entities in the input text. However, the tagger was trained on newswire text, and when applied to blogs and other target domains, it misses a large number of named entities. To increase the recall, we also mark all the phrases that appeared as an anchor for a link in Wikipedia. This step generates a huge number of false positive surface forms. However, in later stages of the Wikifier we use the feature $P(link|s)$ to decide whether to link the surface form s to its highest-ranked disambiguation. The experimental results showed that this strategy was successful at substantially increasing recall with only a minor precision loss.

As mentioned above, we use links in Wikipedia to compute probabilities $P(c)$, $P(c|s)$, and $P(s|c)$ for each surface form s and concept c. The conditional probabilities vary significantly across surface forms. For example, the prior for Michael Jordan (the basketball player) given the surface form "Michael Jordan" is over 99%, while the probability of Tour de France given the surface form "Tour de France" is at around 20% since Wikipedia has a separate page for each year the Tour was held, and the links often link directly to the specific year. However, the general year-free page is still the most common link. To counter this variability, we add the following features: the rank of concept in the list of disambiguation candidates sorted by $P(c_{ij}|s_i)$ and the normalized conditional probability, where the conditional probability of the concept is divided by the conditional probability of the most likely concept (1 both for Michael Jordan and Tour de France). We also use a string similarity metric between the concept title as it appears in Wikipedia and the surface form.

Our Wikifier computes a score for each disambiguation candidate as a linear combination of its feature values. Threshold and coefficient values for each linear combination are optimized using labeled training data. Many optimization strategies are possible; we chose to use a straightforward approach based on stochastic hill-climbing. Starting with an initially random selection of weights, we iteratively randomly perturb these, keeping the new weights if they offer higher performance (in F1) and sometimes otherwise.

We use Explicit Semantic Analysis (ESA) (Gabrilovich & Markovitch, 2009) to mark candidate surface forms. ESA is a technique for binding general and prominent Wikipedia concepts to free text. We use the 20 top Wikipedia concepts extracted by ESA as disambiguation context. The intuition is that disambiguation candidates that appear in the ESA summary are more likely to be relevant, and the corresponding lexical terms are more likely to be linked.

The degree of relatedness of a pair of concepts is appraised by counting the number of other Wikipedia articles referring to both members of the pair. We treat the input text as a "bag of concepts," in contrast to the "bag of words" representation common in Information Retrieval. The features are generated with the intuition that for a coherent assignment of concepts to the mentions in the input text, the concepts are somehow related. Finally, we extract the text of the concept, and all the text surrounding the hyperlinks to the concept. Intuitively, expressions with high linkability, low ambiguity, and expressions that were identified by both the NER tagger, and the ESA, and were often used to anchor concepts in Wikipedia are more likely to have a corresponding Wikipedia page.

Our experiments also show that when we use only an NER tool to detect the mentions, our performance for the Wikification task decreases significantly. We conclude that for general-domain Wikification, NER alone is not sufficient for deciding which surface forms to link. A paper detailing our algorithm, experiments, and results is forthcoming. Test webpages are provided for public experimentation with the NER tool[37] and for our follow-on work in Wikipedia Entity Retrieval (WER) -- the task of retrieving documents from a data collection, where the retrieved documents mention a concept described on a given Wikipedia page.[38]

5.4 Library Applications of Named Entity and Identity Resolution Software

5.4.1 Testing and evaluation

The goal of our evaluation was to determine which type of system best fit our requirements. In order to perform this system evaluation, we undertook a set of standard evaluation steps:

1. Select a balanced evaluation set of test material.
2. Chose a precise markup scheme.
3. Create a set of instructions or guidelines for this markup.
4. Manually label the test material by at least two human labelers.

[37] http://l2r.cs.uiuc.edu/~cogcomp/LbjNer.php

[38] The ongoing-development and demonstration webpage for the wikifier functionality extension is http://l2r.cs.uiuc.edu/~cogcomp/demo.php?dkey=198020101

5. Adjudicate over mismatched labels.
6. Finalize the gold standard test set to use as a baseline for testing systems.
7. Measure additional human performance on labeling task to establish lower and upper bounds on task.
8. Run several named entity recognizers over the gold standard test set.
9. Determine the best match for the EMP application

Step one: select a balanced evaluation set of test material

In order to accurately assess the accuracy and coverage of the Named Entity tagger we have selected, and in order to determine if we will use several Named Entity identifiers in a cascaded architecture, we created a baseline tagged set of material for evaluation and system development. This involved selecting a set of test articles covering the following different domains and genres:

1. Wikipedia article on World War II.
2. Art history texts from Gardner's Art through the Ages (an art history textbook).
3. An excerpt from an Illinois State Legislature document.
4. An extract from Apian, The Civil Wars, from the Perseus Digital Library.
5. A news article from Reuters concerning Hurricane Gustav.

Steps two and three: choose a precise markup scheme and create a set of instructions or guidelines for this markup

After reviewing a number of markup schemes, including MUC7[39] and ACE[40], we chose to use the CoNLL collection for this project which since it is the most commonly used today in natural language processing.

CoNLL uses four categories of tags: persons (PER), locations (LOC), organizations (ORG), and names of miscellaneous entities (MISC) that do not belong to the previous three groups for simplicity. We developed a set of tagging instructions (Appendix A) asking individuals to review the five articles listed above, extract key points and ideas from the text so that search retrieval engines will pick up the information, and annotate each named entity identified using these four tags. In addition, reviewers were asked to insert their rationale for selecting a particular tag with each tagged entity. The rationale was used for informational purposes but was not used in the evaluation of human tagger results.

[39] http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html

[40] http://www.itl.nist.gov/iad/mig//tests/ace/

Results were then analyzed using the following evaluation format:
-Full overlap Example:
-Tagger A: [ORG City University of New York]
-Tagger B: [ORG City University of New York]
-Partial overlap of key nouns Example 1:
-Tagger A: [LOC Mesopotamian]
-Tagger B: [LOC Mesopotamian soil]
-Example 2:
-Tagger A: [PER Tutankhamen]'s tomb
-Tagger B: [LOC Tutankhamen's tomb]
-Partial overlap excluding articles Example:
-Tagger A: [ORG the City University of New York]
-Tagger B: [ORG City University of New York]
-Partial overlap over phrases Example:
-Tagger A: [ORG the City University of New York]
-Tagger B: [ORG University of New York]

Precision and recall is computed by determining whether a tagged Named Entity matches the gold standard. However, unlike other evaluations where a match is easy to judge (for example, in picking a part of speech, usually a word is either used as a noun, adjective, or verb with little subtlety), evaluation for named entities is far more complex. The issue of partial matches was handled as covered in Table 3.

	Actual condition	
	Gold standard (gs)	Not gs
shows there is a NE	TP= True positive. Tagger tagged Gold Standard without superfluous letters/words, e.g., “[PER Gardner [name]]’s” or . “[ORG The Axis [gov’t powers]]” rather than “[ORG Axis [gov’t powers]]”	FP=Tagged as GS but not GS. False positive. This includes improper tags of Gold Standard, e.g. “The Axis” rather than “Axis”
shows there is not a NE	FN=False negative, Gold Standard not detected—missed tagging the GS or any variation thereof	TN=True negative. Did not mark because should not have been marked (used in opposition to false positive)

Table 3. Evaluation Methodology.

Steps four, five and six: select and label a balanced evaluation test set

These next three steps follow the standard approach to creating a baseline annotated resource to measure system performance. Following these procedures, test material was manually labeled by at least two human labelers, then adjudication took place over mismatched labels, and thus the gold standard markup was finalized. During the adjudication process, notes were kept on disagreements for future reference in developing and evaluating the different systems. Through this process, each document had baseline, or gold standard, tagging determined by these three individuals. We documented disagreements as part of establishing a strict upper and lower bound for performance; our notes permitted us to return to recurring regularities in problematic cases during later evaluations as discussed in the next section.

Step seven: measure additional human performance on labeling task to establish lower and upper bounds on task

Next, in order to measure the validity of the baseline gold standard, we ran an additional experiment. Five subjects tagged each document, with experimental controls over subjects, documents, and document types for balance.

Results of manual tagging indicate that taggers did a sufficiently good job applying annotation codes. Precision figures ranged from 62-78% while recall ranged from 71-95%. Problems that stood out in this testing are:

1. systematic coding with wrong category of a tag;
2. disagreement on tagging of the article “the”;
3. some generic terms unexpectedly surfacing as significant, such as (in a sample history text) certain dates of battles and treaties; and
4. one text sample having a marginally lower rate of precision, because it was the first article tagged (although we did some reviewing, not everyone went back and corrected their initial work using their new knowledge – i.e., knowledge gained from having tagged that first text sample).

The full results can be seen in Table 4.

	Wikipedia document	Art History document	Illinois Legislature document	Illinois Legislature document (with outlier removed)	Perseus document	Reuters document
tags	156	44	248	246	102	73
precision	62.93%	64.49%	76.43%	76.43%	78.65%	78.04%
recall	86.02%	71.65%	73.96%	84.67%	95.26%	87.43%
F-score	72.69%	67.88%	75.17%	80.34%	86.16%	82.47%

Precision = tp/tp+fp
Recall = tp/tp+fn

Table 4: Human annotation of 5 sample texts.

Step eight: select and run different named entity recognizers

As discussed in Section 5.2, we compared our tagger with ClearForest Gnosis, the Stanford NER tagger to cover rule-based, hybrid, and statistical systems.

Step nine: determine the best match for EMP application through error analysis

NER systems perform differently depending on the techniques used and on the goal of a project. For example, some systems favor precision over recall, that is, they require a high degree of confidence to recognize a name and then to assign a category to that name. Thus, these systems will recognize fewer names overall, but those that are selected are very likely to be of very high quality and accuracy. On the other hand, other systems might favor recall over precision. In this case, more items will be labeled as potentially named entities, and assigned a probably category; however, more are likely to be incorrect. In the field of named entity recognition systems (as in most of information retrieval) these two factors, precision and recall, typically counterbalance each other.

Thus, systems with very high precision generally have low recall (they find only the best material, but they miss a lot); in contrast, those with high recall generally perform with lower precision (i.e. they find a lot, but there is more incorrect material let into a result set).

In addition to difficulties caused by human tagger error or differences in interpretation as noted above, other challenges arose as a result of the genre of the input text. The art history article drawn from Art through the Ages had poor precision and recall because taggers had difficulty interpreting ornate sentence structures ("The interest in the unearthing of lavish third-millennium b.c. Sumerian burials rivaled the public fascination with the 1922 discovery of the Egyptian boy-king Tutankhamen's tomb (see figs. 3-36 to 3-38).") and metaphorical language ("Nothing that emerged from the Mesopotamian soil attracted as much attention as…"). The wikipedia article had poor precision and average recall due to factors such as erroneous tagging of countries serving as actors, or organizations, rather than locations in a sentence; heavy inclusion of articles in marked entities (e.g., The Axis or The Allies); and frequent marking of years (e.g., 1937). The Reuters and Perseus articles had good precisions and recall results. Precision errors in the Perseus article are due to mis-categorization but recall is extremely high because the substance of the article is straightforward. Lastly, the precision and recall percentages for the Illinois State legislature document improved with outliers removed, but there were some issues unique to this domain including the mis-tagging of complex job titles (e.g. Assistant Doorkeeper of the House Wayne Padget) and names in roll call vote sections. In addition, the article contained significantly more entities than other articles, making it difficult for taggers to focus.

Use-case development

Use cases for NER-tagged text were not difficult to envision because research and product managers in the library community have long been interested in intelligent indexing of full text. This desire is now more urgent, given

1. the need to manage non-MARC records, such as Dublin Core, EAD, and publisher metadata, which contain many unstructured text fields;
2. the results from empirical studies of MARC usage, which show a heavy use of the unstructured text in 3xx and 5xx fields;
3. an increasing need to manage databases of full-text records, such as the QuestionPoint knowledge base; and
4. an emerging business need to resolve the identities of proper names extracted from text by associating them with Wikipedia, library authority files, or aggregated identity resolution resources represented by WorldCat Identities or the Virtual International Authority File.

Of course, the questions raised by these resources could not be answered in a relatively short project such as EMP, but our goal was to begin to bridge the gap between computer science research and library needs and suggest directions for productive work in the future.

As a result, our first priorities were to run the NER tool on library data, evaluate the results, retrain the tool, and configure it to run in a variety of software environments that satisfy use cases prompted by longstanding needs in the library community. A secondary priority was to engage with researchers who were working in parallel on a more leading-edge problem: the resolution of identities

5.4.2 Expanded testing

In the final phase of our work for the EMP project, we extended our experiments with tagging and training to include a set of more diverse materials typically managed by digital libraries, including scholarly articles, whose citations are clearly marked in a "References" or "Bibliography" section; full-text summaries of published works that might be obtained from webpages containing publisher metadata and could be incorporated into MARC 5xx fields; and EAD records, which also contain many full-text fields with carefully written, indexable data. Together, these records supply a variety of record types, styles of presentation, and varying amounts of formal markup that permit us to draw conclusions about what problems will be encountered in a real-world test of NER tagging in library-managed resources; where the NER tool can be applied most productively, given the current state of development; what recommendations we can offer to computer-science researchers that would make future versions of NER tagging utilities more usable for library applications. The outcomes of these experiments are described in Section 5.5 of this report.

The successful conduct of these experiments also required us to become adept at executing the most important workflows associated with the NER tagging task: creating training data, testing large numbers of files for different use cases, and scoring the results, which in turn required that we fill the gaps in the published workflows, producing output that could be easily passed from one step to the next. These processes are described below.

Running the NER tool

The UIUC NER tool has a command-line interface, accompanied by instructions for creating a shell script for running the tool to tag a document and produce results marked up in the CoNLL tagging scheme. This installation is sufficient for small-scale experiments, but we made several enhancements to increase usability. First, a newly constructed outer layer permits the tool to run in a greater variety of software environments. In one configuration, the NER tool can be accessed from a Web service API that is appropriate for service-oriented interactive applications, which also required a refactoring of the code to correct security links. In another configuration, which researchers at OCLC considered to be more useful, the NER tool was optimized to process batches of files instead of individual files, which required that a time-consuming initialization step is performed only once per job, not on each file. Finally, the API layer now allows users to specify four different output syntaxes:

1. The normal bracketed CoNLL fomat, such as [PER John Doe];
2. An XML format very similar to CoNLL, such as <PER>John Doe</PER>, which is used by software utilities that can process XML but not CoNLL;
3. An HTML format that assigns different colors to the four name categories, such as <span class='PER'>John Doe</span>, which is used for visualizing the tagged data in the context of the input file; and
4. A line-oriented format that outputs only the tagged text, which is used for tabulating the tagged output.

A copy of the NER tool, the Web service interface, the enhanced batch interface, and the run script, as well as additional technical documentation, are available in the EMP project open source repository (Appendix B).

Training the NER tool

Out of the box, the NER tool is trained on newswire text, which is sufficient to demonstrate proof of concept and can give usable results on similar kinds of texts. But given the range of text genres and use cases encountered by applications in the library community, developers need to be adept at training the tool to suit their needs. Training requires that a human expert create so-called 'gold' data by marking up a corpus of plain text with the CoNLL encodings of the correct tag assignments, from which the NER tool induces patterns that are labeled with the four categories of names: [PER], personal names; [LOC], location names; [ORG], organizational names; or [MISC], the names of miscellaneous human-created artifacts. For most texts, this task requires that HTML, XML, or structural other markup be stripped out first.

Given that a corpus size of 500 K or more is required for adequate training, the task can be tedious and error-prone. Essentially the same task must be performed when evaluating the accuracy of the NER tagger, which requires that the output from the NER tagger be compared with correctly tagged output. We distinguish between 'training' gold and 'scoring' gold to make it clear that two workflows are involved, but the gold data produced is exactly the same: plain text with CoNLL tags that identify four kinds of names.

To aid in the creation of gold standard data, we developed a configurable HTML and XML stripping tool that permits users to extract all of the text from a marked-up document, or only from segments enclosed in specified tags. A more significant contribution is a visualization tool that permits users to view the tagged text as color-coded HTML, XML, or CoNLL markup -- formats that are supported by the redesigned NER tagger output layer described above. Through a point-and-click interface, users can cancel a tag, change the name of the tag, or tag a new span of word tokens in the input text. This utility speeds up the creation of gold data and eliminates a potentially catastrophic error that is easy to commit when the task is done by hand: the failure to balance the opening and closing brackets of a CoNLL tag. Unbalanced gold data can still be processed by the NER tool, but it gives unexpected results and, in some circumstances, can crash or hang the system.

A copy of the training script for the NER tool, the HTML stripper, the visualization utility, sample gold data, and associated technical documentation are available in the Subversion repository (Appendix B).

Scoring the results

Once the NER tool has been trained and new output has been generated, the results can be evaluated by comparing the automatically tagged text with the scoring gold. Precision and recall and an F score representing a ratio of the two measures[41] are calculated by the scoring tool developed by the University of Maryland and described in more detail in Section 5.4.1 of this report.

We made several enhancements to the scoring tool. First, results are displayed in a Web browser instead of an Excel spreadsheet. Second, scores can be reported for a set of files instead of a single file. Finally, separate F scores can be calculated for each tag type, which makes it possible to assess the accuracy of the name types across multiple files and permit researchers to draw conclusions about the kinds of tags that can be accurately extracted from texts representing different genres or subject domains. Scores from the generalized reporting tool are reported for the experiments described in the next subsection.

The enhanced scoring tool and associated technical documentation are available in the Subversion repository (Appendix B).

Creating workflows

As the previous discussion implies, running and training the NER tool and evaluating the results involves many steps. We have filled in gaps, automating as many steps as possible because we learned that lack of context is a major barrier to the deployment of utilities such as the NER tool by users who are not experts in natural language processing or machine learning. Accordingly, the Subversion repository (Appendix B) has all of the material required to execute the most important workflows, with detailed instructions, sample data, run scripts, and pointers to the required programs and utilities. The following workflows are documented:

1. Prepare texts for tagging or training.
2. Create gold data, either training gold or scoring gold.
3. Run the NER tool in training mode.
4. Run the NER tool in tagging mode.
5. Score the results.

[41] `F = (P*R)/(P+R))*2`

Government documents

Identifying functional genre

Provisions for collecting or archiving digital documents, including here the evaluation of the NER tool, can be informed by knowledge of the genre of documents likely to be encountered. Although different aspects of collection and curation may classify documents into genre based on differing criteria (e.g., size, file format, subject), this research is primarily interested in classification based on the functional role the document plays in state government, akin to (Toms, 2001), but specifically utilizing documents from Illinois State Government (ISG). The classifications listed herein are based on an overview of ISG digital documents, encountered in over nine years of gathering and archiving work with and for the Illinois State Library (ISL), and on discussions with practitioners in cataloging and in government documents librarianship. The associated technical report (Jackson, 2010b) provides genre definitions and numerous (currently) Web-accessible example documents.

State government documents are interesting in this research in that they are presumably somewhat comparable to both federal government documents and business documents. Perhaps surprisingly, there are also portions of the State Web that are somewhat less than businesslike, either in tone or in technological proficiency of implementation. In this respect state government digital documents may also be useful approximations to documents produced either personally or by small activities. Having a list of government document genre can also inform work in information promulgation (e.g., through website design, or the design of a series of printed materials), and the grouping of documents for digital library or archival purposes.

In the case of ISG documents, developing a genre classification can be assisted through analysis of the archived copies and online digital document collections developed by ISL, beginning in 2001[42] (Jackson, 2003 & 2005). Instead of just speculating on the nature of digital publishing that might be going on, these extensive collections were inspected. Accordingly, this work included a review of the 330 Illinois State Government websites currently online, inspecting homepages and the top-most levels to identify collections of document-like information products.

[42] This work includes website archiving (i.e., the PEP IMLS National Leadership Grant project at http://www.cyberdriveillinois.com/departments/library/who_we_are/pep.html and the CEP IMLS National Leadership Grant project at http://www.cyberdriveillinois.com/departments/library/who_we_are/cep.html), a permanent digital library for official publications (Electronic Documents of Illinois - EDI - at http://ediillinois.org/), a central collection facility for community-contributed, scanned-to-digital material (Illinois Digital Archive - IDA - at http://www.idaillinois.org/) and a search engine encompassing all ISG websites (Illinois Government Information - IGI - at http://igi.finditillinois.org/cgi-bin/search.cgi). Jackson's research team designed, constructed, and for multiple years operated PEP, CEP, EDI, and IGI. They also operated CEP for six other states, each for at least a year.

These functional genre were suggested as a result of working with the sources listed;

1. Legislation: Legislation necessarily, and uniquely, has the form necessary for the authoritative formal statement of the text of laws and regulations. Some may wish to include sub-genre here, such as amendments and other follow-on material. Legislation is often enacted to honor some person or group, or to document some event in the public record. Legislation documents generally include measures to facilitate the operation of the legislature (e.g., line numbers).
2. Requirements, Codes, Regulations, and Laws: This material, too, is impacted by a need to be a formal statement of a requirement, code, regulation, or law. These formats are for purposes other than recording the business of the Legislature and the associated facilitation of editing or amending source text.
3. Oversight Reports: Oversight reports are recurring reports, addressing the performance of a significant portion of the reporting agency's role within state government (e.g., and quintessentially, annual reports). Reports are typically addressed, however accurately, to a specific body having oversight responsibility, but perhaps using the tone of a report to the wider citizenry or to the legislature. Topics addressed center around normal and emergent special activities of the year, for at least a substantial portion of the resources of the agency authoring the report.
4. Special Topical Reports: Special reports are typically; (a) topically focused (e.g., scientific, agricultural, fiscal, assessments), (b) not recurring, or at least aperiodic, and (c) are motivated by events or special circumstances, which may continue or recur. They may provide topically-focused information in some depth, as opposed to addressing in detail a large portion of the efforts of the agency authoring the report.
5. Newsletters and Periodicals: Newsletters are typically; (a) an incremental source of information, (b) usually of a timely nature, (c) address only a small part of the total activities of the authoring agency, and (d) are typically intended for readers already familiar with the major activities of the agency or the major topic being addressed.
6. Informatory or Introductory Material: Informatory or introductory material would, in print, typically be a flyer or single-page handout. In-depth information would not be addressed, except possibly in extreme summary. Webpages of this nature may exist to help readers navigate to

the section of the website appropriate to their specific information need, for example, simply introducing the existence of an agency or parts thereof.

7. Instructional Material: Instructional material includes specific information, in sufficient depth useful for the performance of tasks or the fulfilling of responsibilities. The size of these documents is typically a few pages, at most. Not included here (but, as forms and instructions, below) are instructions related to the filling in of a form. Also not included here are educational materials.
8. Slides: Informational material, perhaps in useful depth or occasionally specific, but formatted for slide-based presentation, presumably involving an instructor.
9. Budgetary Material: Budgetary material addresses budget and fiscal matters at the State or State agency level. Smaller scale discussion of fiscal matters generally does not perturb the purpose or structure of the entire document.
10. Audits: Audits are perhaps recurring, although also perhaps aperiodic and event driven. They are written by the agency conducting the audit. They may address fiscal matters, although alternatively or additionally may address an appraisal of the audited agency's performance of its duties.
11. Legal Proceedings: Records of the actions, or applications to a court or oversight board.
12. Minutes: Minutes are formally recorded for a wide variety of government meetings and activities (e.g., formal meetings, town hall meetings, hearings, court proceedings other than transcripts, appeals).
13. Plans or Projections: Plans or projections are primarily narrative, although frequently encompassing some fiscal material and imagery, etc., for various illustrative purposes.
14. Two-Dimensional Displays: Two-dimensional displays include blueprints, maps or other GIS information products (e.g., aerial photography), and charts or graphs. It may also include non-GIS photography, if such is an information product of a government agency, or otherwise involved in the agency discharging its function (e.g., items in a museum exhibit). Obtaining blocks of text from within maps to support name recognition processing may be highly problematic.
15. Contractual Material: Contractual material contains the legal language and records of doing business with the State. This includes the administration of grants of various types.

16. Memoranda: Memoranda, particularly memoranda of understanding, generally address the clarification or elaboration of policies, or the coordination of intra-State activities.
17. Forms and Instructions: Forms and instructions address the conducting of the huge number of ways individuals and businesses interact with the State. Most use a form of form, existing online and/or in a paper version.
18. Kids' Material: Unexpectedly, and possibly re-election related, government officials frequently feel the need to produce "kid-friendly" information by-products. Kids' material is generally educational and youth-oriented. Publications are intended for younger readers, and are generally written at a corresponding reading level.
    Other demographic groups could be included in a broader genre definition here (e.g., under a label “material targeted for specific demographic groups”), if the writing style or genre changes specifically for that group, but such practice does not appear to be the case, aside from the obvious use of non-English languages for some groups. Subject headings are defined for material of interest to specific demographic groups[43], and agencies or major activities within agencies address specific issues related to such groups[44], but the nature of the writing style and choice of publication format does not appear to differ from the norms for government publication (e.g., simplified language and considerable use of cartoon-like imagery is done for kids’ publications, but there is no analog for publications of particular interest to age/racial/income/locale/social groups).
19. Directories: A directory provides a list of people, places, or organizations. If large, these may be produced by computer program, and if so, pagination may differ markedly between implementations.
20. Website Locator and Navigation Webpages: Website locators and navigation webpages are generally informal, or are tersely menu-like. They generally don't contain much specific information, but instead link to webpages/websites where additional material is available. These are "hubs" [Ingwersen, 1998], pointing to information sources.

[43] For example, “Social issues and programs: Ethnic groups and minorities: American Indians” and “Laws and regulations: State statutes: Laws concerning the elderly”.

[44] For example, the Illinois Department on Aging, the Illinois Assistive Technology Project, the “Parenting 24/7” and “IllinoisParents“ websites, the African-American Family Commission, the Deaf and Hard of Hearing Commission, and the Illinois Early Childhood Collaboration.

21. Social Media and Interactive Communication Facilities: Social media communication mechanisms (e.g., blogs, YouTube videos, Flickr photos, Facebook discussions, Ustream, MySpace, Twitter) are incorporated in or linked from several Illinois state agency websites. Some of these sponsoring agencies, or the associated preservation agencies, are attempting to preserve this information content. For some contexts, this may constitute a genre.
    Two practical matters make archiving this material problematic; (1) spider-based harvesting from these host systems can be highly problematic due to the wide variety of ways system vendors may employ scripts and databases in the storage and presentation of the material (i.e., in-depth cooperation by the website/facility operator would seemingly be needed), and (2) some of the discussion facilities are not owned/operated by ISG, and may claim ownership of content posted thereon.
    The question of whether these materials are "government documents" and/or appropriately archived is moot in the Illinois case in that ISL does not consider this material within their charter. ISL has bypassed archiving this material thus far.
22. Press Releases: Press releases, perhaps called news releases or briefings, typically exhibit a different, third-person writing style. These releases are very frequently issued, and are correspondingly numerous.
23. Datasets: Datasets are necessarily gathered in connection with all manner of studies, but access to the raw data via the Web appears to not yet be occurring for Illinois. An IGI search for "dataset" produced only 22 hits across all Illinois State Government websites, and the preponderance of those either were reports mentioning the dataset used, or documents specifying how a certain dataset is to be collected (i.e., the awarding of funding for data-gathering, and enumerating the fields to be filled in). Apparently only one Illinois website provides direct access to data (following), and even so this data is only presented in Statewide totals. Drill-down is not supported, probably out of privacy-based necessity.
    ISL considers collecting or archiving datasets/databases outside their charter, so such acquisitions have not knowingly been done. (Statement of data records in brief, formatted, text-like webpages, though, can result in the contents of some database being downloaded and archived inadvertently.)

Further, tabular data may simply not contain natural language narrative, making its use with tools such as examined in the EMP project moot. Accordingly, datasets are not further analyzed as a genre in this study.

24. Information Facilitating Recreation: Some government publishing addresses the potentially recreational use of facilities under cognizance of the government (e.g., parks and public lands, or genealogical use of records). In cases other than Illinois', where such publishing is more extensive, it may be desirable to consider this a distinct functional genre. In the Illinois case, seemingly sections 3.5, 3.6, 3.16 and 3.17 cover documents which could serve the purpose of facilitating recreation. Accordingly, such documents are not pursued further in this study.
25. State Academic Institutions: State governments support academic institutions of multiple types, to varying extents. The purposes of academic institutions are so different from those of other state agencies, and perhaps so unique in the types of information they must manage (e.g., facilities for and records of student interactions, faculty biographies, coursework resources, publications resulting from research, and descriptions of the social scene) as to separate them from the existing genre. Certainly they would need to be archived differently (e.g., academic institutions often have institutional repositories of their own). As the colleges and universities of Illinois are only partially funded by the State, and as ISL has adopted a policy of not attempting to archive or index the contents of academic institution websites any deeper than the upper-most few levels, the various document genre originating within academic institutions are set aside in this study.
26. Newspapers of Public Record: States utilize newspapers where official legal notices are occasionally or periodically published. ISL does not archive any digital versions of these, but instead archives the digital material arising from the Legislature. The content some states may print in a newspaper of record appears to be, in the case of Illinois, addressed in items 1, 2, 9 to 13, 15, 19, and 22, and as such is not examined further.
27. Correspondence: Examination of correspondence to or from government officials was not pursued in this study due to privacy concerns and lack of document availability on the Web.

Sampling and dataset preparation

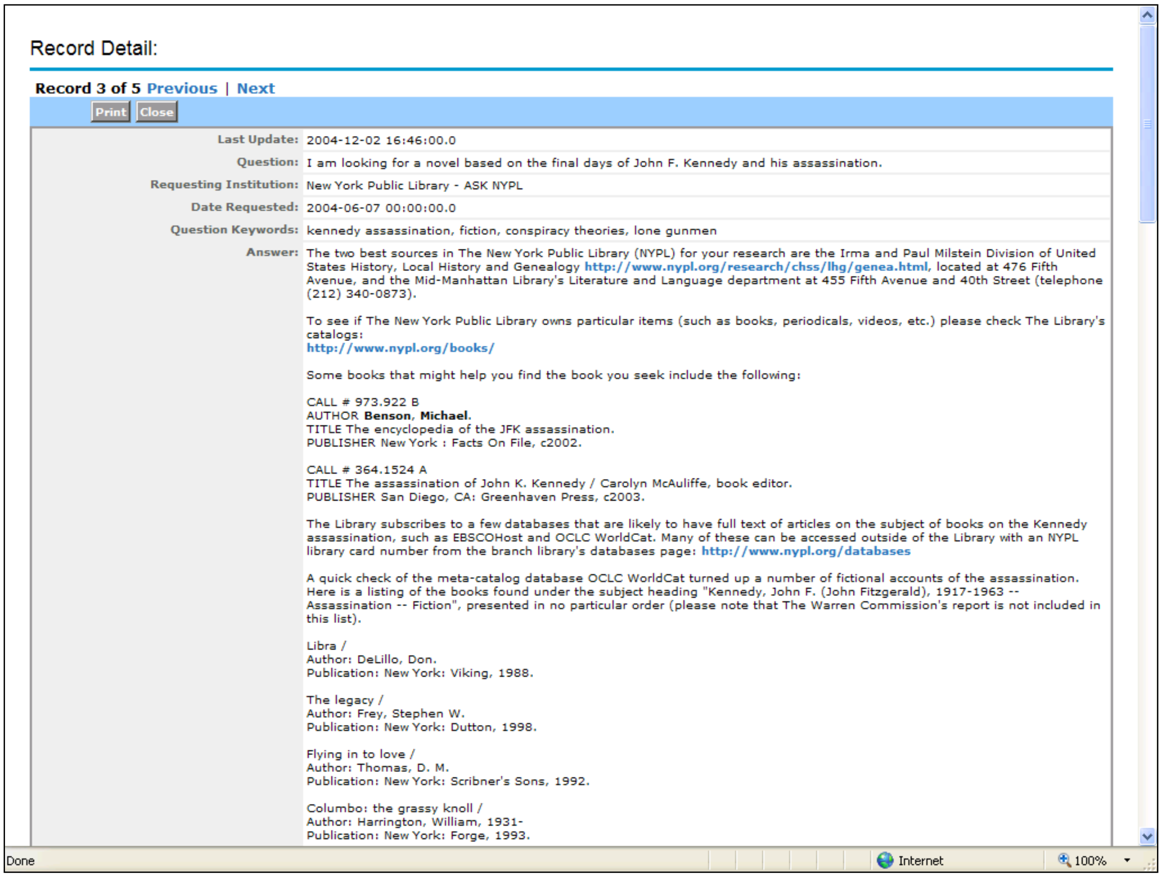
Lists of examples for each document genre were developed, encompassing 30,791 documents. These lists were then randomly sampled, three examples per genre, and the conversion of the various file formats to plain text was done. Conversion to text used (a) "save as" functionality built in to various document format editors or viewers, (b), copying and pasting from viewer displays into text-only editor utilities (e.g., Notepad, vi), and (c) the UNIX 'ps2txt' software, capable of converting PDF as well as PostScript format. These text files were then the input to the named entity recognizer tool, in a lengthy exploratory testing series performed by OCLC.

QuestionPoint

OCLC's interest in the technology developed in the EMP project stems from the need to link unstructured text to its large collections of highly coded records, such as bibliographic and authority records, and other metadata required to support the management and discovery of library resources. OCLC researchers are now turning their attention to the many streams of full text that are associated with these materials, such as author biographies, reader reviews, online reference works, unpublished or pre-published manuscripts collected in institutional repositories, and similar materials. In the terminology developed in the problem statement above, the association of unstructured text to structured metadata is necessary, because the coded material often has the identities, while the unstructured text is what mentions one form of the name.

Consider an example from QuestionPoint, the virtual reference service maintained by OCLC in partnership with the Library of Congress.[45] Library patrons submit a question through the QuestionPoint interface, which is automatically routed to the closest participating librarian, based on the IP address of the computer from which the question originates. The librarian answers the question in a response window after a time delay that varies from a few minutes to a few days. Questions and answers that are of general interest are eventually collected in a database, which users can search and browse. Figure 8 shows one example of a question-answer record, which is a full-text document. If readers want to find out more about the broad topic, John Fitzgerald Kennedy, or the authors of the books cited in the librarian's answer, they may associate this record to other resources at OCLC and elsewhere. But they would have to cut and paste selected text into WorldCat.org, Google, Wikipedia, or other resources that might provide more depth or context. The interface doesn't do this work for them. In other words, this text frequently mentions names, but identity resolution is up to the reader.

[45] http://www.oclc.org/services/brochures/211401usb_questionpoint.pdf



Last Update:	2004-12-02 16:46:00.0
Question:	I am looking for a novel based on the final days of John F. Kennedy and his assassination.
Requesting Institution:	New York Public Library - ASK NYPL
Date Requested:	2004-06-07 00:00:00.0
Question Keywords:	kennedy assassination, fiction, conspiracy theories, lone gunmen

Answer:

The two best sources in The New York Public Library (NYPL) for your research are the Irma and Paul Milstein Division of United States History, Local History and Genealogy **http://www.nypl.org/research/chss/lhg/genea.html**, located at 476 Fifth Avenue, and the Mid-Manhattan Library's Literature and Language department at 455 Fifth Avenue and 40th Street (telephone (212) 340-0873).

To see if The New York Public Library owns particular items (such as books, periodicals, videos, etc.) please check The Library's catalogs:
http://www.nypl.org/books/

Some books that might help you find the book you seek include the following:

CALL # 973.922 B
AUTHOR **Benson**, **Michael**.
TITLE The encyclopedia of the JFK assassination.
PUBLISHER New York : Facts On File, c2002.

CALL # 364.1524 A
TITLE The assassination of John K. Kennedy / Carolyn McAuliffe, book editor.
PUBLISHER San Diego, CA: Greenhaven Press, c2003.

The Library subscribes to a few databases that are likely to have full text of articles on the subject of books on the Kennedy assassination, such as EBSCOHost and OCLC WorldCat. Many of these can be accessed outside of the Library with an NYPL library card number from the branch library's databases page: **http://www.nypl.org/databases**

A quick check of the meta-catalog database OCLC WorldCat turned up a number of fictional accounts of the assassination. Here is a listing of the books found under the subject heading "Kennedy, John F. (John Fitzgerald), 1917-1963 -- Assassination -- Fiction", presented in no particular order (please note that The Warren Commission's report is not included in this list).

Libra /
Author: DeLillo, Don.
Publication: New York: Viking, 1988.

The legacy /
Author: Frey, Stephen W.
Publication: New York: Dutton, 1998.

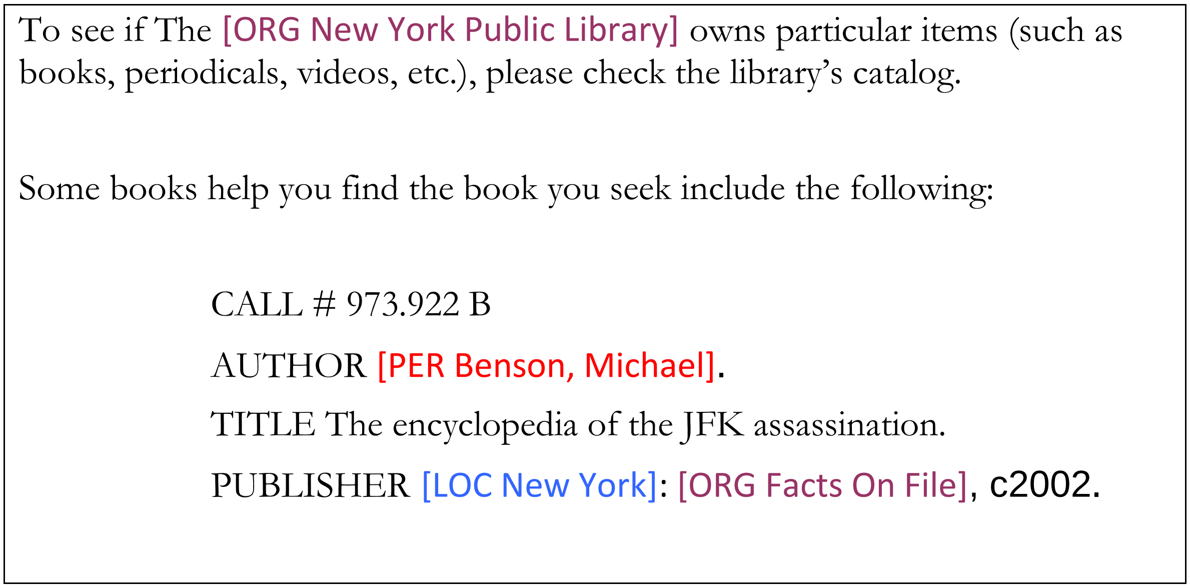
Flying in to love /
Author: Thomas, D. M.
Publication: New York: Scribner's Sons, 1992.

Columbo: the grassy knoll /
Author: Harrington, William, 1931-
Publication: New York: Forge, 1993.

Figure 8. A record in the QuestionPoint knowledge base.

With more sophisticated information extraction from unstructured text and algorithms that link the output to structured resources, the records in this database could be enhanced to add clickable links to the QuestionPoint record. When these links become available, the reader would, with minimal effort, be able to find The Encyclopedia of the JFK Assassination or The Assassination of John F. Kennedy in his/her local library, find a list of other books by the author Michael Benson or the editor Carolyn McAuliffe (listed in Figure 8), and discover other works about the assassination of John Fitzgerald Kennedy, or related broader and narrower topics. Since the structured metadata already supports such exploration, the only missing piece is the association with texts such as the QuestionPoint answer. The EMP tools are designed to provide this information.

The first step is to run the QuestionPoint record through the LBJ Named Entity Tagger to obtain the name occurrences. The results are shown in Figure 9. Organizational names are green, locations are blue, personal names are bright red, and miscellaneous names are brownish red.



To see if The [ORG New York Public Library] owns particular items (such as books, periodicals, videos, etc.), please check the library's catalog.

Some books help you find the book you seek include the following:

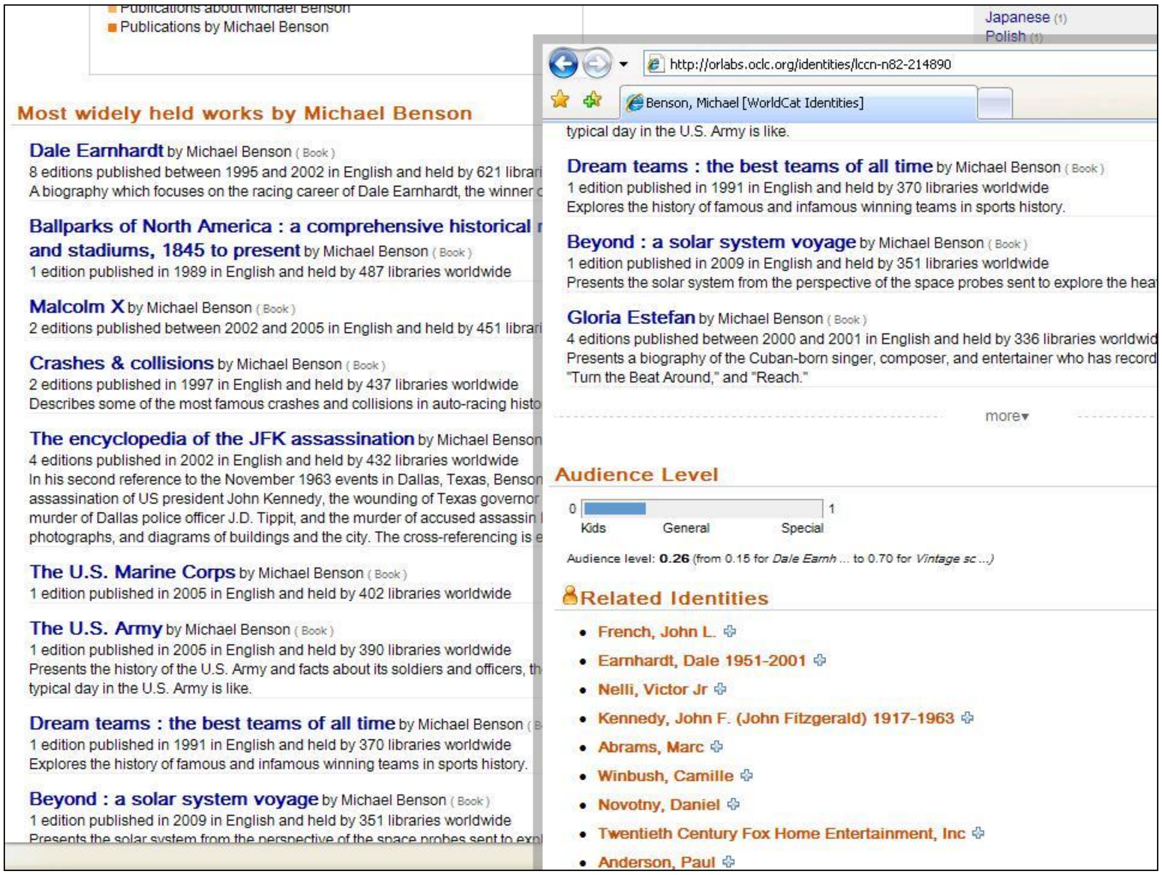
CALL # 973.922 B
AUTHOR [PER Benson, Michael].
TITLE The encyclopedia of the JFK assassination.
PUBLISHER [LOC New York]: [ORG Facts On File], c2002.

Figure 9. NER markup for a fragment of a QuestionPoint answer.

In initial tests with the QuestionPoint answer records, the most important problem is the parsing and linking of the book citations, shown here. To obtain useful output from the NER tool, we had to overcome some built-in bias and train it to recognize names of the form [PER Last, First] and [PER Last, First Initial]. With about 450K of training data, we obtained results that recognized these new forms while retaining the tool's native ability to recognize names conforming to the more usual [PER First Last] pattern. The training data also specifies that any name following the pattern [LOC] and a colon (:) is an organization, leading to the correct recognition of publisher names. The title remains untagged, but it is recognized through a regular-expression match as the text that intervenes between the pattern [PER Last, First] and [LOC]:[ORG], as shown.

Once the name occurrences have been extracted and selected, the next step is to link them to the correct identities. The obvious tool for accomplishing this goal is the Wikipedia tool being developed by Ratinov and Roth, which enables linking the name occurrences to Wikipedia, but this turns out not to be useful. Although Wikipedia has an entry for Michael Benson, the name annotated in Figure 9 above, it describes the documentary filmmaker, not the author of The Encyclopedia of the JFK Assassination, the title annotated above in Figure 9. The deeper problem is that Wikipedia is not the best identity resolution authority for the task of assigning clickable links to book citations, because it contains relatively few articles about authors.

WorldCat Identities is a more promising authority. The page for Michael Benson, the author of The Encyclopedia of the JFK Assassination, is shown in Figure 10. This page has a rich collection of links for this author, including a list of his published books, alternative forms of the author's name, a list of co-authors (with indirect links to their published works), and a list of subject headings associated with the author.



- Publications about Michael Benson
- Publications by Michael Benson

Most widely held works by Michael Benson

Dale Earnhardt by Michael Benson (Book)
8 editions published between 1995 and 2002 in English and held by 621 librari
A biography which focuses on the racing career of Dale Earnhardt, the winner c

Ballparks of North America : a comprehensive historical r
and stadiums, 1845 to present by Michael Benson (Book)
1 edition published in 1989 in English and held by 487 libraries worldwide

Malcolm X by Michael Benson (Book)
2 editions published between 2002 and 2005 in English and held by 451 librari

Crashes & collisions by Michael Benson (Book)
2 editions published in 1997 in English and held by 437 libraries worldwide
Describes some of the most famous crashes and collisions in auto-racing histo

The encyclopedia of the JFK assassination by Michael Benson
4 editions published in 2002 in English and held by 432 libraries worldwide
In his second reference to the November 1963 events in Dallas, Texas, Benson
assassination of US president John Kennedy, the wounding of Texas governor
murder of Dallas police officer J.D. Tippit, and the murder of accused assassin
photographs, and diagrams of buildings and the city. The cross-referencing is e

The U.S. Marine Corps by Michael Benson (Book)
1 edition published in 2005 in English and held by 402 libraries worldwide

The U.S. Army by Michael Benson (Book)
1 edition published in 2005 in English and held by 390 libraries worldwide
Presents the history of the U.S. Army and facts about its soldiers and officers, th
typical day in the U.S. Army is like.

Dream teams : the best teams of all time by Michael Benson (B
1 edition published in 1991 in English and held by 370 libraries worldwide
Explores the history of famous and infamous winning teams in sports history.

Beyond : a solar system voyage by Michael Benson (Book)
1 edition published in 2009 in English and held by 351 libraries worldwide
Presents the solar system from the perspective of the space probes sent to exp

Japanese (1)
Polish (1)

http://orlabs.oclc.org/identities/lccn-n82-214890

Benson, Michael [WorldCat Identities]

typical day in the U.S. Army is like.

Dream teams : the best teams of all time by Michael Benson (Book)
1 edition published in 1991 in English and held by 370 libraries worldwide
Explores the history of famous and infamous winning teams in sports history.

Beyond : a solar system voyage by Michael Benson (Book)
1 edition published in 2009 in English and held by 351 libraries worldwide
Presents the solar system from the perspective of the space probes sent to explore the hea

Gloria Estefan by Michael Benson (Book)
4 editions published between 2000 and 2001 in English and held by 336 libraries worldwid
Presents a biography of the Cuban-born singer, composer, and entertainer who has record
"Turn the Beat Around," and "Reach."

more▾

Audience Level

Audience level: **0.26** (from 0.15 for *Dale Earnh* ... to 0.70 for *Vintage sc* ...)

Related Identities

- French, John L.
- Earnhardt, Dale 1951-2001
- Nelli, Victor Jr
- Kennedy, John F. (John Fitzgerald) 1917-1963
- Abrams, Marc
- Winbush, Camille
- Novotny, Daniel
- Twentieth Century Fox Home Entertainment, Inc
- Anderson, Paul

Figure 10. The WorldCat Identities page for Michael Benson.

WorldCat Identities is created algorithmically, primarily by collecting data from OCLC's WorldCat database. Preprocessing utilities mine WorldCat's bibliographic records, creating a separate page for every author, as well as for every person (real or fictitious) who has been the subject of a published work. But in a database the size of WorldCat, there are many authors named Michael Benson. How does the algorithm link to the correct author?

The answer turns out to be elegantly simple. The key insight is that the name of the author and the title of the book can be thought of as a bigram, in which the first element is Michael Benson and the second is Encyclopedia of the JFK Assassination. Significantly, an author-title bigram is highly improbable and often unique. In other words, it is unlikely that more than one Michael Benson authored a book with this title about the JFK assassination. Since WorldCat Identities can be searched from an API that accepts an Open URL, a publicly accessible specification for representing information typically found in a bibliographic record (Van de Sompel & Beit-Marie, 2001), the author and title can be sent in the form shown here:

```
http://worldcat.org/identities/find?url_ver=Z39.88-
2004&rft_val_fmt=info:ofi/fmt:kev:mtx:identity&rft.namelast
=Benson&rft.namefirst=Michael&rft.title=MICHAEL+BENSON+AND+
THE+ENCYCLOPEDIA+OF+THE+JFK+ASSASSINATION+%28+%27.
```

This URL triggers a fuzzy-name search against WorldCat Identities, which returns a results list containing a list of 49 Michael Bensons. The top-ranked Benson, the correct link, goes to the Identities pages shown above in Figure 10. To finish the task of presenting clickable answers to QuestionPoint queries, a software routine embeds this intelligence into the XML of the text that is served through the user interface.

This example shows that in a best-case scenario, the problem of associating book citations found in full text with a link that disambiguates the author's name can reduce to the problem of name recognition. Once the name occurrences have been correctly extracted from the input text, sophisticated search and ranking algorithms already in place generate the candidate identities and recommend the correct one.

Other problems at OCLC involving links between resources resemble the QuestionPoint example, but it is instructive to make the underlying issues more explicit. In the example we have discussed, the name occurrence is in the unstructured text and the identity is in a collection of structured resources, which constitute an identity resolution authority. There may be more than one identity resolution authority, which may have complementary strengths. The task of disambiguating the name of a book author is best accomplished by referring to an identity resolution authority that is customized for the published record. However, if the task is to establish the identities of names of local historical or cultural figures, about whom little or nothing has been published, Wikipedia may be a better authority than WorldCat Identities. These observations imply that identity resolution algorithms will perform better when multiple resources can be consulted. It is a priority for future work to determine how this is best accomplished.

Yet a more significant issue emerges from this data. What happens when no available name resolution authority can resolve a name occurrence? A name would still be extracted from unstructured text, along with other identifying characteristics, such as a book title, if the name is an author; birth and death dates, if the person is famous; a subject domain associated with the person's work, and so on. But if no match can be made even against a detailed text, the text itself now contains one form of name occurrence as well as important clues for resolving the identity. If these clues are collected, they could form a valuable first draft for a larger and more timely identity resolution resource that is populated automatically, a huge improvement over the current state of the art.

Testing on IMLS collection-level metadata

In addition to NER applications in whole text processing, name extraction will also have application in query standardization. Queries may be run against bibliographic databases, and as such questions arise of the effectiveness of an NER tool when applied to metadata.

The online form of the NER tool[46] was tested (Jackson, 2010a) using 31 of the 293 collection-level descriptive metadata files OAI-harvested from the Grainger Engineering Library's federation of IMLS-sponsored digital collections under the IMLS Digital Collections and Content "Opening History" project[47].

The collection records used in this test were those also being used as test data by the Collection-Item Metadata Relationships (CIMR) project at GSLIS (Renear, Wickett, Urban, Dubin & Shreeves, 2008). RDF-formatted metadata was obtained from the CIMR project, who in turn obtained it by OAI-harvesting the metadata from Grainger. Item-level metadata could not be used as the most promising narrative-like field ('`description`') is occupied by URLs for item-level metadata records. These URLs might be thought to start the reader on a quest which will ultimately disclose an item description.

Portions of this metadata was originally written by the owners/operators of the various digital collections. Other portions, particularly those addressing collections as a whole, reflect standardization work or implementation decisions done in the process of metadata federation. Only the collection metadata tag named `dcterms:abstract` and `dc:title` were utilized here, in expectation their contents would be largely based on natural language, and as such might reasonably conform to the design assumptions of the NER tool.

The abstracts of collection-level metadata are generally quite narrative in character, and as such fit the design assumptions for whole-text processing quite well. Error rates and types did not seem appreciably different than in EMP project tests using the NER tool. Titles, however, engendered problems in correct delineation of field values. Initial-upper-case characters, when encountered in the words of a title, are no longer reliable indicators for NER. Further, variance in case usage within titles federated from multiple sources is also possible.

[46] http://l2r.cs.uiuc.edu/~cogcomp/LbjNer.php

[47] http://imlsdcc.grainger.uiuc.edu/history/.

The EMP NER tool can benefit, particularly in metadata application contexts, from expanded training data including:

1. capitalization practices typical of titles,
2. by implication, completely-capitalized strings (e.g., such as may be imported from databases), and
3. street addresses.

Other fields were only briefly examined as being either unlikely to contain text which included names (e.g., dates, URIs), or which probably contained only a name (e.g., publisher identification), enclosed between the XML tags, making name recognition basically a foregone conclusion. However, it is clear from the results obtained that the EMP project default training data did not address street/mailing addresses much, as the initial numbers are generally omitted from the recognized location (LOC) substrings. So, there was no point in belaboring that issue.

How various words somewhat synonymous to, or associated with "collection" in this context[48] are to be labeled is also not a foregone conclusion. In one sense, these might be the names of an organization -- certainly they could plausibly occur on the wall of the associated building, or in the sponsoring organization's letterhead or homepage masthead. But, a collection of artifacts is not the same thing as the organization collecting those artifacts, nor is the project funding the collecting.

5.5 Conclusions and Recommendations

The QuestionPoint exercise is a simple proof-of-concept demonstration for a set of processes that start with the automatic extraction of names from unstructured input text and end with significant enhancements to a commercially available product. This is a work in progress, however. The most immediate need is for improved recognition of the large variety of book and article citation styles in text that was not designed for machine processing. Similar problems are being addressed by other researchers at OCLC who are using the NER tool to extract names from text fields in a bibliographic record, with the goal of increasing the navigable links in collections of published works.[49]

In fact, there is no shortage of uses for robust NER extraction and identity resolution utilities in the library community. A name extractor tool can also be used to parse names that occur in collections of digitized government documents. But it will have to be expanded to recognize not only the names of persons, locations, and

[48] e.g., "Gayle Morrison Files on Southeast Asian Refugees," "Augustus F. Hawkins Papers," " John Wesley Powell Expeditions," and "William Edward Hook Glass Negatives."

[49] See, for example, the demo of Work Records In WorldCat, accessible at: http://frbr.oclc.org/research/pages/026336461.html. The field named 'Derived Terms' was populated with the LBJ-NER tool.

organizations, but also government information applications, position titles, edifices, geographic features, geo-political regions, and laws or regulations. Once found, these names can provide searchers many more precise access points into collections than are currently available through state-of-the-art systems.

Key persons, places, concepts, and artifacts occur in information retrieval in almost all disciplines, making progress on the identity resolution problem a broad, cross-cutting need. Outside library circles, identity resolution authorities would need to be created from scratch. Large numbers of topically focused communities have literatures emerging on the Web, thanks in no small part to prototypes and best practices developed under IMLS and Library of Congress research funding.

In the next, post-NDIIPP-2, phase of development we will address the disambiguation of recognized names resulting from such software. We plan to run our named entity extraction software on a variety of directory-like webpages[50] as a means of facilitating the initial construction of name authority files, with an eye to establishing "community-authored" authority lists. The University of Illinois has done extensive work in archiving digital or digitized state government documents, resulting in a vast collection of materials. Notoriously rich in name variations, these digital government materials would support this stage of investigation extremely well. Efficient citizen (and government staffer) access into that corpus would benefit considerably from name disambiguation.

5.5.1 Discussion and recommendations

Tagging

The UIUC NER tool extracts and named entities into four categories: locations (LOC), persons (PER), organizations (ORG), and miscellaneous names of human-created artifacts (MISC). But the LBJ-NER tool does not permit nested tags, which means that users must make a choice when confronted with the task of assigning labels to names such as 'New York Times.' If the application requires that location names have priority, the string should be labeled as [LOC New York] Times. If it requires that organization names have priority, the string should be labeled as [ORG New York Times]. To capture the information required for both applications requires the assignment of nested tags such as [ORG [LOC New York] Times], which is beyond the scope of the current generation of NER taggers.

The NER tool also limited because it works only on the literal text that is submitted as input. But a human reader can make a simple inference from text strings such as 'Translated by Jacques and Jean Duvernet' – namely, that 'Duvernet' is the surname

[50] Official webpage lists and rosters like the State of Illinois Telephone Directory (http://www.illinois.gov/teledirectory/printable.cfm) published by the Governor's office, or the Illinois General Assembly's list of Illinois State Senators (http://www.ilga.gov/senate/) are available, although socially contributed Web lists such as Wikipedia pages may similarly be processed.

of both Jacques and Jean because it appears inside an English phrase structure that linguist have dubbed 'Conjunction Reduction,' which eliminates redundancy in phrases linked by a conjunction. But the NER tool can only tag the text it encounters, producing the output 'Translated by [PER Jacques] and [PER Jean Duvernet].' A dedicated application that calls the NER tool and applies some awareness of English syntax would be required to generate the more complete tag [PER Jacques Duvernet] and [PER Jean Duvernet]. Similarly, subsequent mentions of names in a given text are often represented in reduced form such as 'Mr. Duvernet,' 'Duvernet,' 'he,' or 'the translator,' which refer back to a named entity using a pronoun or alternative phrasing, a grammatical construct called anaphora. Without an anaphora-aware process that would track the mentions and associate them with the appropriate named entity, the NER tool can only discover and tag those mentions that exhibit some features of personal names, such as a capitalized first letter or a word that is recognizable as personal-name vocabulary such as 'Jacques' or 'Anderson.'

Another limitation is the inescapable ambiguity of natural language, which cannot be detected or corrected by an automated tagging tool whose purpose is to classify text strings by skimming text for superficial cues. The result is that, even in a model with only four categories of names, the human expert who creates the gold data must read the text carefully and make a strategic decision about how to tag nearly every chunk of text in the corpus and apply those decisions consistently throughout a potentially large corpus, an error-prone task.

For example, does the string 'Currier & Ives' contain the names of two people, which would require the tags [PER Currier] & [PER Ives]? Or is it an organization that publishes the work made famous by the two artists it is named after, which would require the tag [ORG Currier & Ives]? Or are the tags essentially arbitrary because the text is so vague that the distinction doesn't matter (e.g., "Illinois highways")? Similarly, should 'White House' be tagged with [LOC] because it is a well-known place, [ORG] because the term is used metaphorically to refer to the executive branch of the United States Government, or [MISC] because it is the name of a human-created artifact? Only a close reading of the text, coupled with a clear sense of how the tags will eventually be used, can provide guidance. Despite this pervasive problem, distinct readings can be confidently assigned in a small number of cases because they appear in contexts that are regular and abundant enough to be detected by a machine-learning algorithm. For example, when the string 'H.H. Abrams' appears in the context of a citation such as "TREASURES OF THE AMERICAN ARTS AND CRAFTS MOVEMENT, 1890-1920] by Tod M. Volpe, Beth Cathers, and Alastair Duncan] (New York: H.N. Abrams, 1988)," it can be tagged not as [PER], but as the eponymous name of an organization. Names of publishers that are not ambiguous between the [PER] and [ORG] reading, such as "Random House," are detected in the same context.

These limitations must be considered as the state of the art in NER tagging is evaluated for its potential to move up the chain from intriguing examples, to research prototypes and demos, and eventually to software applications that solve tangible problems for users in the library community. The sections below describe some of the issues associated with each class of named entities recognized by the LBJ-NER tagger. To shed additional light on tagging performance, results were obtained for a sample of QuestionPoint records, EAD records, and government documents using the UIUC NER tool trained on news text. They were scored using the enhanced scoring tool available from the project repository. The records were chosen because they help us understand what might happen when NER tagging is deployed in the management of documents typically found in a digital library.

Personal names

NER tagging results on personal names show that the tool is mature and useful for indexing on unstructured and semi-structured text, with F-scores ranging from 60-81%. In OCLC's experiments, we were most interested in the appearance of the personal name in the context of an article or book citation, where several patterns of personal names are observed. Training was required to recognize names conforming to the patterns {Last name, First name} and {Last name, First initial}, which was accomplished without breaking the NER tool's innate ability to recognize personal patterns of the form {First name, Last name} or name strings consisting of a single token. With a sufficient supply of consistent training examples, the NER tool was induced to make fine distinctions in the markup of personal names such as those illustrated in Table 5 (below). In the first example, the name string 'Taylor, Joshua C.' ends in the middle initial, whose period is correctly included inside the right bracket. In the second example, the name string 'Craven, Wayne' consists of a last name followed by a complete first name; the trailing period is thus part of the citation style, not the name.

[PER Taylor , Joshua C.] THE FINE ARTS IN AMERICA ([LOC Chicago] : [ORG The University of Chicago Press] , 1979)
[PER Craven , Wayne]. AMERICAN ART HISTORY AND CULTURE (New York: [ORG H. N. Abrams] , 1994)

Table 5. Examples of problematic personal name markup.

Because of the extra attention given to training on documents containing citations, the F-scores for personal names were highest on the QuestionPoint data and slightly lower on government documents and EAD records.

In future work, a dedicated personal-name application could extend the functionality of the NER tool by extracting names from conjoined and anaphoric expressions. The NER tool itself could also be made more useful by increasing the granularity of the name tagging, by assigning separate tags to name components such as {First name}, {Last name}, {Title}, or {Honorific}. The ONIX standard [ref], which is widely used in the publishing community, or the emerging International Standard Names Identifier (or ISNI) standard[51], which is proposed for name and reputation management in the academic community, can provide lightweight guidelines that inform decisions about which categories should be tagged.

Location names

Precision and recall results are high for the NER tagging of location names and the outputs are useful for indexing. Two issues were raised in our experiments on library data. First, the definition is broad because it includes geographic features such as lakes and mountains as well as features in the human built environment, which raises questions about ambiguity mentioned above. In other words, is 'New York Public Library' a location or an organization? The same question would not have to be asked about Mount Rainier or the Rhine river. Second, aggressive tagging of locations introduces the problem of excess specificity. For example, since street addresses are locations, the NER tagger converts strings such as '476 Fifth Avenue' or '455 Fifth Avenue and 40th Street' into tagged output, though it is questionable whether [LOC Fifth Avenue] is accurate, [LOC 476 Fifth Avenue] is useful for any application, or whether the intersection of Fifth Avenue and 40th Street implied in the second string should be lost when the tagger assigns separate [LOC] tags to the two names.

Organizational names

Precision and recall results are lower for the NER tagging of organizational names, with F-scores ranging from 25-45%. As with location names, the definition is broad because it doesn't distinguish between corporate names ([ORG Google]), government entities ([ORG Division of Wildlife]) or ([ORG Iraqi Army]), and ephemeral gatherings of individuals ([ORG Democratic Party Meet-up]).

In addition, organization names can be quite long, and unlike personal or location names, there is no expectation that they have any predictable compositional

[51] http://www.isni.org/

structure. This is a consequence of the fact that an organizational name is for a human-designed construct, which can be anything involving a collection of people that is worth talking about and can assume any imaginable form whatsoever that succeeds in establishing a reference. As Downing insightfully argued about English compound nouns over thirty years ago (1975) , the creator of a name is under no obligation to follow a template, use particular words, or devise anything that is clever, concise, or even memorable. Unfortunately, this insight is at odds with what a named entity recognition tool is attempting to accomplish because it needs some sort of guidance to define a category—which it obtains either from a built-in vocabulary, a set of built-in heuristics, or from detectable patterns in a corpus. When the built-in assumptions are violated, the NER tagger does badly, as in this string, 'Irma and Paul Milstein Division of United States History, Local History and Genealogy,' a name for a division of the New York Public Library. The NER tagger erroneously tags it with the sequence '[MISC Irma] and [PER Paul Milstein] Division of [LOC United States History], Local History and Genealogy,' instead of assigning a single tag that encompasses the entire string: [ORG Irma and Paul Milstein Division of United States History, Local History and Genealogy]. It is not obvious that training data can alleviate the problem because the name does not follow a repeatedly observed pattern.

Related to this problem is that, in unstructured or lightly edited text, the writer is not obligated to use the full legal name of the organization or any established variants. As a result, the text may contain an idiosyncratic descriptor such as 'Natural History museum' instead of 'American Museum of Natural History.' Such descriptors are normally considered to be outside the scope of the named entity recognition task because they reflect the natural productivity of language and may be infinite, while names are established, relatively frozen, and countable. Nevertheless, the NER tagger cannot make this subtle theoretical distinction, so it tags the descriptor as [ORG Natural History museum] instead of classifying it as a non-name.

Miscellaneous names

The [MISC] tag was not useful in OCLC's analysis because the definition is too broad. Perhaps because of this problem, F-scores were extremely low – less than 10% for all classes of text. What do 'World War I,' 'Abstract Expressionism,' 'Kleenex,' 'The Proper Treatment of Quantification in English,' and 'Jewish-American' – all tagged with [MISC] in various tests with the NER tool have in common, aside from the fact that they are proper names instead of ordinary words and phrases, but are not names of people, places, or organizations? In our experiments, we treated [MISC] as an open category that could be trained to assume a narrower definition. As described in the previous section, the most successful experiment used [MISC] to tag article and book titles.

Recommendations

As the comments above imply, named entity tagging is a complex psycholinguistic task that challenges even mature, sophisticated readers. Thus the tagging task can only be approximated with a model that recognizes just three broadly-defined categories, plus a fourth category with limited utility, none of which can be assigned any internal structure. In the short term, researchers in the library-science domain who wish to apply this technology must lower their expectations, define tasks that can be carried out successfully with the current state of the art, and identify productive avenues for future enhancements. Below are some recommendations motivated by our work.

1. The most successful and mature categories are already valuable for indexing and some types of data mining. Several projects at OCLC are already underway to incorporate the information they provide. Nevertheless, it would be helpful to be able to narrow or otherwise refine the definitions, either by making the installation configurable (through the addition or deletion of gazetteer entries), or by supplying or suggesting training data that could accomplish the goal.
2. Personal names and names of locations have predictable internal structures that can be decomposed and modeled, which an enhanced named-entity tagger should be able to recognize. Without access to the name's internal structure, the most obvious applications in the library community are those that require only flat strings, such as displayable browse indexes.
3. The LBJ-NER tool is based on perceptrons, which use minimal assumptions to assign names and make multiple passes through the text to acquire evidence that may change the initial category assignments. This is an excellent solution to the problem of NER tagging in library data because it can be configured to handle many genres, rhetorical styles, and subject domains.

   In OCLC's limited experiments with the [MISC] tag, we were able to use training data to induce a narrower definition than was originally implemented. Such a solution could be extended to the of names of laws, government documents, song titles, webpages, and other specialized names that appear in full-text documents managed by libraries that are not now being indexed and may not exhibit the problems with metaphorical speech noted in the prior discussion that limit the application of named-entity tagging. But with just four categories, the output of the NER tool is still severely limited. A valuable enhancement would be the inclusion of empty categories in the NER tool that users could define with training data and additional gazetteer entries, if necessary.

4. Applications of the NER tool on library data are more successful when the input is edited or consists of semi-structured text in a limited subject domain. If the text is too variable, training data can't be assembled to contain enough evidence for some names of interest. But if the text is highly structured and carefully edited, users need to consider whether their goals could be accomplished with a less sophisticated tool than an NER tagger based on machine-learning algorithms.
5. Training the NER tool is tedious, error-prone, and often intellectually challenging. For some applications, such as the experiment we conducted with EAD records, it is difficult to assemble enough training data from the corpus of interest, so a corpus with similar characteristics must be used instead. Since the need to create training data poses a significant barrier to the widespread deployment of machine-learning-based technologies such as NER tagging, care must be taken to ensure that the time investment is strategically targeted. We identified several opportunities for future work in this report, including the tagging of authors and titles in lists of citations and open-ended discussion, or organization names in government documents. We concluded that the development of training data can only be productive if: 1) the text contains many proper names of interest to librarianship and exemplify learnable patterns; 2) criteria for applying the markup can be articulated and consistently applied to the data by human experts; 3) the recommended markup falls within the scope of the tagging scheme produced by the NER tagger; 4) the patterns cannot be easily discovered by simpler means, such as regular-expression matching; and 5) the corpus containing the patterns is large enough to change the behavior of the NER tool.
6. It is unlikely that sophisticated applications can be built using advanced technologies such as machine-learning-based NER tagging unless they are deployed in the context of other advanced technologies that can classify segments of text or perform at least rudimentary natural language processing tasks such as extracting keywords and phrases or resolving inferences such as those produced by conjunction reduction. In other words, there is limited utility in applying the NER tool as a black box.

5.5.2 Identity resolution

Identity resolution is necessarily a less mature endeavor in this project because the research record is still unstable. Nevertheless, section 5.4.2 of this report describes OCLC's interest in identity resolution in the context of the QuestionPoint knowledge base, which contains many mentions of works that are catalogued in OCLC's WorldCat database. Right now, the user can access them only by cutting and pasting the relevant text into a browser search box. In other words, the identity resolution problem on QuestionPoint records is easy and the output would improve an existing product, resulting in a visible payoff for the effort invested.

Unfortunately, the article and book citations embedded in unstructured text in the QuestionPoint records are difficult to discover automatically. Sometimes they appear in a labeled list of references and sometimes they appear in a text with a full reference, but often they are simply given a passing mention—all of which adds to the variability of the citation text and makes automated recognition difficult. With a success rate hovering below 50%, it was not realistic to develop a demo or a mature research prototype until we made progress on this problem. But this is a problem with tagging, not identity resolution.

We tried to solve this problem by using the [MISC] tag more strategically to identify the title of the cited text. Though this solution had limited success, other approaches to the problem of identifying cited books and articles in unstructured text might turn out to be more fruitful. For example, a machine learning algorithm might be trained to perform a binary classification of text streams either as citations or as non-citations, using such clues as the presence of the pattern [PER Last name, First name] or a sequence of capitalized tokens set off by periods. As a last resort, we can recommend that the editors mark the citations that need to be linked, as is routinely done in scholarly articles that end with a clearly marked bibliography or list of references. Note that these suggestions for making further progress would implement three of the recommendations listed in the previous section, namely:

1. NER tagging is most effective in a software context that includes other sophisticated algorithms;
2. Focus on a small subset of discoverable patterns; and
3. Work with edited text.

To check the accuracy of named entity recognition on texts that do not present a citation recognition problem, we performed a preliminary test with a small number of research articles from DLib Magazine. Though many more of the citations can be recognized, identity resolution is more challenging in this context. Some citations, such as "Duval, E., Hodgins, W., Sutton, S. &amp; Weibel, S.L. (2002). Metadata principles and practicalities. D-Lib Magazine 8(4)" have exact matches in WorldCat and a corresponding entry in WorldCat Identities. Others such as "Chan, Sebastian. OpenCalais meets our museum collection / auto-tagging and semantic parsing of collection data. Fresh + New(er). Powerhouse Museum. March 31, 2008" have indirect matches -- which means that the author is represented in WorldCat Identities but the work is not. But this is valuable information because it presents an opportunity to enhance an author's Identities page and is, in fact, OCLC's primary motivation for processing this data. If links to article titles, blog posts, or grey literature by or about an established author can be created, WorldCat Identities can evolve into a rich and constantly evolving resource for navigating the published record. But even more value can be added by mining the information contained in formal citation lists or informal mentions of authors in running text. For example, citations such as "Anderson, James D. and Melissa A. Hofmann. 2006. A Fully Faceted Syntax for Library of Congress Subject Headings. Cataloging & Classification Quarterly 43, 1: 7-38." have no match at all in WorldCat Identities. Yet this context contains important information for establishing an identity -- such as co-authors, dates, venues of publication, or author-supplied keywords found in the citing text, which identify a subject domain for the work. All of these clues could be used to generate a rough draft for an authority record that might act as stand-in or a collection point until it can be edited by human experts. Obviously, we are early in our investigation of these issues, and our work will continue after the EMP project has been completed.

We are also monitoring progress on the UIUC Wikifier, described in section 5.3.1 of this report. Though our applications are tailor-made for identity resolution authorities focused on the published record, there are many ways that our work could intersect with the Wikification research, which we expect to pursue after the EMP project is concluded. For example, the Wikifier can be used to assign identities for names that are not well-established in the published record, either because the authors have not yet produced traditionally published works or because they work in segments of society where this is not likely ever to happen. Secondly, and perhaps more simply, WorldCat Identities itself represents a corpus of text that can be wikified, so the UIUC Wikifier could be applied as a utility to associate every possible Wikipedia page with appropriate WorldCat Identities pages. Finally, since the two identity resolution resources are built with fundamentally different technology stacks, it is possible that both resources can be made more robust by incorporating the algorithms used in their counterpart.

6. INFORM Risk Assessment Methodology Project

6.1 Introduction

The importance of standards and best practices in the creation and sustainability of digital content is a given in the digital library community. Standards and best practices act as a common language, enabling the interoperability of systems, and hence promoting the exchange of information and ideas. This is no less so for the digital preservation community, particularly where the preservation risk assessment of digital materials is concerned.

Significant effort has already been invested by the digital preservation community in the creation of resources such as the Global Digital Format Registry, PRONOM, JHOVE and Droid. These resources have been intended to provide a shared infrastructure to enable identification of file formats used in the construction of digital objects and to maintain repositories of representation information regarding those materials. While this information allows repository managers to know what, exactly, they hold in their repository and to have some certainty that representation information exists to enable the decoding of information in selected formats, this is an extremely partial view of file formats and does not provide managers with the information they need to conduct the most fundamental preservation analysis of a file format: is it safe to use?

In order to address this lack, staff at OCLC's Digital Collections and Preservation Services division proposed the development of a new methodology, INFORM. The INFORM methodology applies traditional risk assessment tools to the problem of evaluating the durability of digital file formats for preservation purposes. Such a methodology, through the application of a common metric for the evaluation of file formats, should allow multiple organizations to engage in risk assessment of digital file formats and contribute them to a common pooled resource of assessment information. Through this resource, format specifications could be compared, essentially providing a "metrics to communicate the measurements [of preservation durability] to a wide audience, recognizing differences in awareness, expertise, language, and interests" (Stanescu, 2004). These are the concepts behind the INFORM Risk Assessment Methodology Project (also called the INFORM project in this report).

In this final report, documenting INFORM project activities, Sections 6.2 through 6.4 explain the methodology for risk assessment in greater detail; this methodology serves as the basis of the software application developed on the project. Sections 6.5 through 6.9 describe the development of the tool (including an explanation of its interface and functionalities); the design of our research protocol; and the outcomes of our data-gathering and analysis.

6.2 The INFORM Methodology

The goals of the INFORM methodology are to investigate and measure risk factors of digital file formats; provide guidelines for preservation planning; and objectively analyze risk trends. These are accomplished by defining the following:

- Risk factors of digital formats and their dependencies
- Risk categories for each format
- Scales to measure probability of occurrence and impact
- Methods to collect, interpret, and report the results

The methodology also classifies risk, as exemplified by Table 6 below. A risk class is essentially a characteristic of the file format that must be assessed for preservation durability.

Classes of risk	Definition of risk class
Digital object format	Risk introduced by the format specification itself and by dependent specifications of compression algorithms, DRM, encryption, etc.
Software	Risk introduced by all necessary software components such as operating systems, applications, library dependencies, migration programs, etc.
Hardware	Risks introduced by necessary hardware components, including CPU, I/O cards and peripherals
Media	Risks introduced by necessary media associated with the format
Associated Organizations	Risks related to the organizations supporting in some fashion the classes identified above

Table 6 – Classes of risk listed and defined

6.3 The Risk Assessment Model

The INFORM tool operates on this risk assessment model:

Risk exposure =
(probability of an accident producing a loss) x (impact [or size] of the loss)

For the classes of risk listed above, in Section 6.2, a scale for risk assessment—estimating the probability that a hazard will occur on a data format—was developed (Table 7), with **1** denoting very low risk and **5** denoting very high risk.

Very low	Below 1%	1
Low	Between 1% – 5%	2
Moderate	Between 6% - 10%	3
High	Between 11% - 25%	4
Very High	Above 26%	5

Table 7 – Scale for assessing probability of hazard on data format

For these same risk classes, an impact (size of loss) assessment scale was created (Table 8), with **A** denoting minor, or insignificant, data loss and **E** denoting catastrophic, or unavoidably complete, data loss.

Minor	Insignificant	A
Significant	Preventable Partial Data Loss	B
Serious	Preventable Complete Data Loss	C
Very Serious	Unavoidable Partial Data Loss	D
Catastrophic	Unavoidable Complete Data Loss	E

Table 8 – Scale for assessing loss impact for data format

6.4 The Risk Exposure Result

6.4.1 Combining probability and impact

The risk exposure result, calculated as the product of hazard (loss) probability and loss impact, is tabled below (Table 9). Note the legend underneath the table, which describes the actions to take, based on the calculation of risk exposure.

E	5	6	7	8	9
D	7	8	11	12	4
C	5	6	9	10	3
B	3	4	3	4	2
A	1	2	1	2	1
	1	**2**	**3**	**4**	**5**

Table 9 – Risk Exposure Result table
Legend: *light gray = watch, gray = prepare, black = act*

The result of any evaluation is a triple showing summed assessments for each zone. (Any class of risk may have individual factor assessments falling in different zones.) For instance, given a risk class with two assessments of **1A,** then an assessment of **3A** and **3D,** and another of **2E,** a triple such as the following would result:

- (2x1A, 3A+3D, 2E) =
- (2x1, 1+11, 6) =
- (2, 12, 6) where **2 = watch result, 12 = prepare result, and 6 = act result**

6.4.2 Figuring in dependencies

Dependencies should be considered in the determination of preservation risk. Any given format, hardware, software or media may depend on organizations responsible for maintenance and creation. In addition, formats may depend on formats, hardware on hardware (with media included in the hardware), and software on software.

For each separate class, examine all dependencies' triple scores and use a MAX function to produce a combined result, illustrated in the example below:

1. File Format Assessment = {13, 5, 3}
2. Associated Organization = {9, 12, 0}
3. Combined Risk Assessment =
    - MAX ({13, 5, 3}, {9, 12, 0}) =
    - **{13, 12, 3}** → *larger value*

In a situation where a given format has dependencies on hardware (including media) and software, one would compute the triple scores for format, hardware/media, and software classes associated with the format and then average the results. This example is shown below:

1. Format = {13, 5, 3}
2. Software = {14, 14, 6}
3. Hardware/Media = {8, 1, 0}
    - AVG ({13, 5, 3}, {14, 14, 6}, {8, 1, 0}) =
    - **{12, 7, 3}** → *average of results*

6.5 Overview of the INFORM Project

The sections above lay out the methodology that guided the functionality of the INFORM tool, developed on this project. Work on the INFORM project occurred over three phases during ECHO DEP 2:

- Phase 1: Software development
- Phase 2: Research protocol development
- Phase 3: Data gathering and analysis

The key objective in developing the software application was to make the methodology easy to access and use, as well as to share and process the results. During the research protocol development stage, we identified formats to assess and groups to target for undertaking preservation risk assessment. (Phase 1 and Phase 2 began in parallel, with Phase 2 ending by the close of the second quarter of 2008; Phase 1 was carried out over all of 2008.) The data-gathering and analysis of Phase 3 (which began at the start of 2009 and concluded in 2010) essentially constituted the pilot part of the project, in which test users were recruited to try out the tool and to report back—via follow-up, semi-structured interviews with the project leads—on its usability and its potential role in their local digital preservation workflow.

6.6 Phase 1 – Software Development

6.6.1 Start-up tasks

The software development activities that took place early in Phase 1 were mainly start-up tasks. A test server was established, with development and testing tools, a CVS version control system, and backup systems; a database model centered on the Global Digital Format Registry UML model, which was created and implemented in MySQL, was set up; and test data drawn from the Digital Cinema System specification were created. Throughout this phase, code for implementing the actual survey process was completed.

Software development for the complete INFORM tool continued for most of the project's first year, as an assessment model schema was created; and code for outputting information from the database into the model schema was developed. These activities necessitated conferring with the Global Digital Format Registry [GDFR], which became the Universal Digital Format Registry [UDFR] upon merging with PRONOM in August 2009. However, during the course of our project, OCLC—one of the key participants in developing the GDFR—withdrew from further work on the registry, returning control of future system development to Harvard University. With respect to the GDFR, Harvard's attention had been focused primarily on the transition process and simplifying the installation process for the GDFR software. While OCLC's withdrawal did not present any insurmountable problems for the INFORM project, it did mean that the project became a lower priority for OCLC. As a result, some of the planning and design work of the tool progressed more slowly than originally had been anticipated.

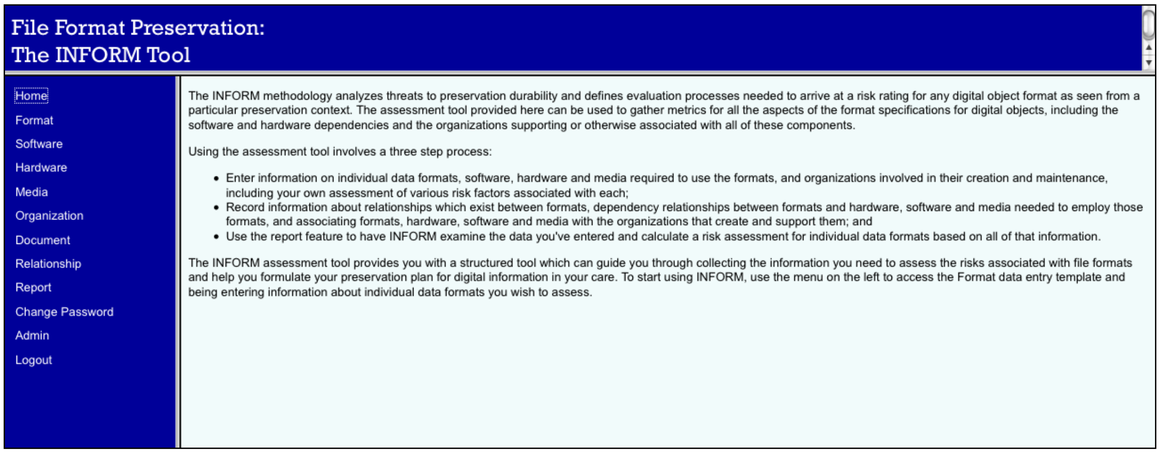
During Phase 1, OAI-PMH support was added to the system. Compliance of the INFORM database with the OAI-PMH would mean that INFORM metadata records, in particular those conveying the risk assessment associated with the format being described, would be harvestable by repositories.

In addition, the team decided upon the formats to be studied, or analyzed: evaluation focused specifically on video and audio formats in use for preservation purposes within the public television system and within major research institutions that have digital preservation repositories.

During the initial stages of the project, it became clear there are relatively significant issues regarding the comparability of risk assessment data produced at different institutions. We surmised fairly quickly that determining the effectiveness of INFORM as a collaborative tool would require a qualitative research approach, enabling us to understand the ways in which local contexts influence determinations of risk. Thus, shifting gears slightly, we decided to pursue a more qualitative method to data collection and analysis. See Section 6.7 for details on the research protocol we designed for more in-depth data gathering, as well as Section 6.8 for a report on the data and its analysis.

6.6.2 The user interface of the INFORM tool

Also during Phase 1, the project team designed a user interface for survey implementation, which was modified iteratively, as warranted by feedback from test users, over the course of the pilot project phase. The home page of the INFORM tool is shown below (Figure 11).



The INFORM methodology analyzes threats to preservation durability and defines evaluation processes needed to arrive at a risk rating for any digital object format as seen from a particular preservation context. The assessment tool provided here can be used to gather metrics for all the aspects of the format specifications for digital objects, including the software and hardware dependencies and the organizations supporting or otherwise associated with all of these components.

Using the assessment tool involves a three step process:

- Enter information on individual data formats, software, hardware and media required to use the formats, and organizations involved in their creation and maintenance, including your own assessment of various risk factors associated with each;
- Record information about relationships which exist between formats, dependency relationships between formats and hardware, software and media needed to employ those formats, and associating formats, hardware, software and media with the organizations that create and support them; and
- Use the report feature to have INFORM examine the data you've entered and calculate a risk assessment for individual data formats based on all of that information.

The INFORM assessment tool provides you with a structured tool which can guide you through collecting the information you need to assess the risks associated with file formats and help you formulate your preservation plan for digital information in your care. To start using INFORM, use the menu on the left to access the Format data entry template and being entering information about individual data formats you wish to assess.

Figure 11 – Home page for the INFORM tool (http://sulfite.lis.uiuc.edu:8080/InformProject/).
Note: the tool requires a login and password.

The navigational sidebar displays the classes of risk, introduced in Section 6.2, that need to be assessed:

- Format
- Software
- Hardware
- Media
- Organization

In addition to the above features, there are links for Document and for Relationship. The Document feature allows the user to input information about a new document, such as updated specs for a format. The Relationship feature captures information about relationships that may exist between formats and about dependency relationships (such as between formats, software, and hardware).

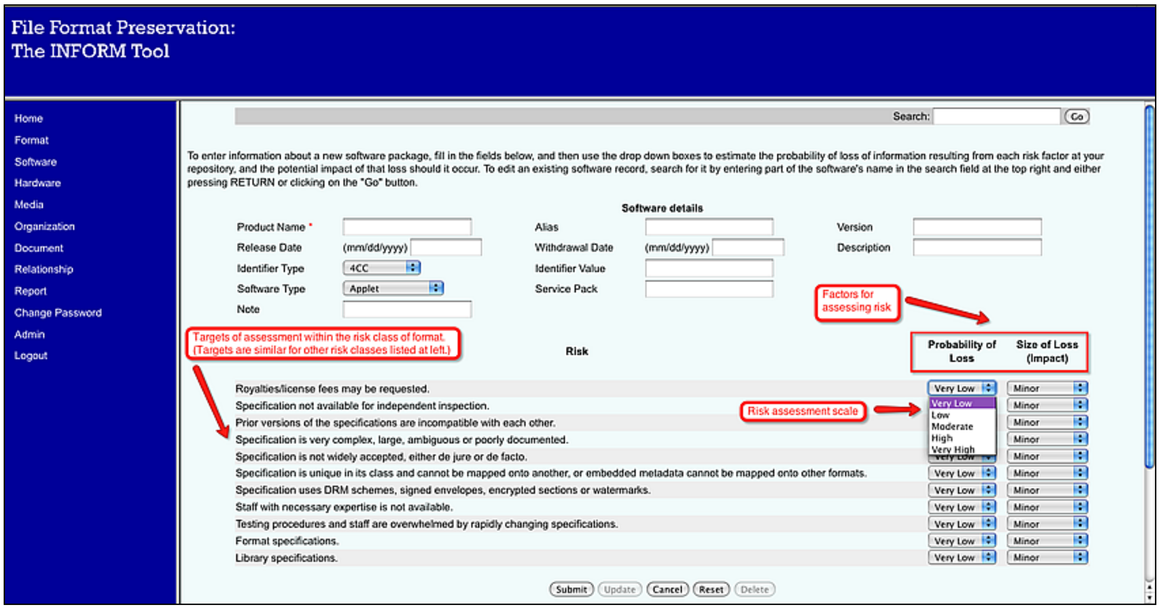
Use of the INFORM tool involves a three-step process (explained on the home page of the site):

1. The user enters information on individual data formats, software, hardware, and media (required to use the formats), as well as organizations involved in their creation and maintenance, including local assessment of various risk factors associated with each;
2. Then information is recorded about relationships existing between formats, dependency relationships between formats and hardware, software, and media needed to employ those formats, and associating formats, hardware, software and media with the organizations that create and support them; and
3. The report feature is activated, enabling the INFORM tool to examine the data that has been entered and to calculate a risk assessment for individual data formats based on all of input information.

Figure 12 shows the Format page, which is one of the pages where Step 1 occurs. Here, basic information ("Software Details") about the format is requested. Pull-down menus at right allow users to select risk assessment on a scale of Very Low to Very High (for probability of loss) and of Minor to Catastrophic (for impact). Users are asked to rate probability of loss and size of loss in regard to the specification of the format and to dependent specifications of compression algorithms, DRM, encryption, etc. The Format page is similar in layout and content to the pages for Software, Hardware, and Media.

Also, at the bottom of each page (underneath the buttons for "Submit," "Update," etc.) are more detailed versions of the pull-down menus depicting the range of risk assessment. These definitions (figure 13) are like those of the tables seen in Section 6.3 above.

Once information for the format and each class of risk has been entered, the user may apply the “Report” feature to generate a report: the INFORM tool examines the data entered by the user and calculates risk assessment for the individual data formats based on the information from the user. Figure 14 contains a sample reports for the Digital Cinema Specifications System.



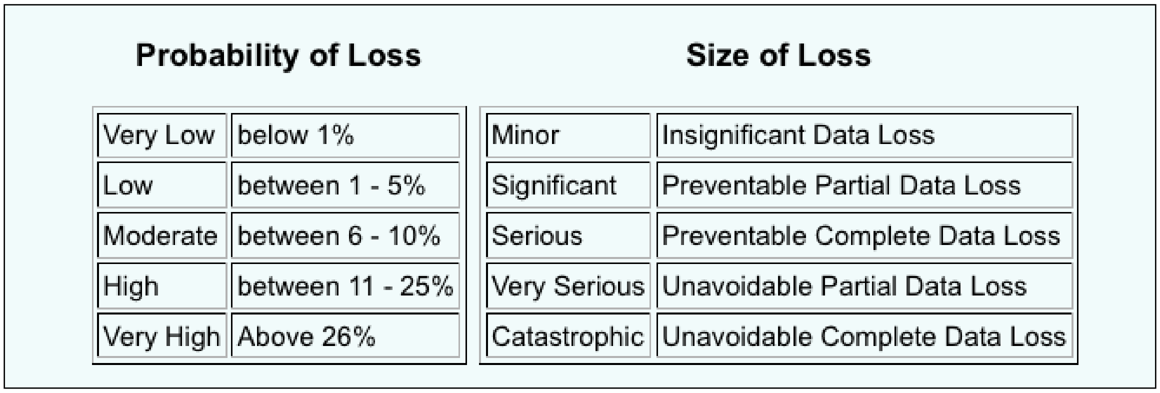
To enter information about a new software package, fill in the fields below, and then use the drop down boxes to estimate the probability of loss of information resulting from each risk factor at your repository, and the potential impact of that loss should it occur. To edit an existing software record, search for it by entering part of the software's name in the search field at the top right and either pressing RETURN or clicking on the "Go" button.

Software details

Product Name *		Alias		Version	
Release Date	(mm/dd/yyyy)	Withdrawal Date	(mm/dd/yyyy)	Description	
Identifier Type	4CC	Identifier Value			
Software Type	Applet	Service Pack			
Note					

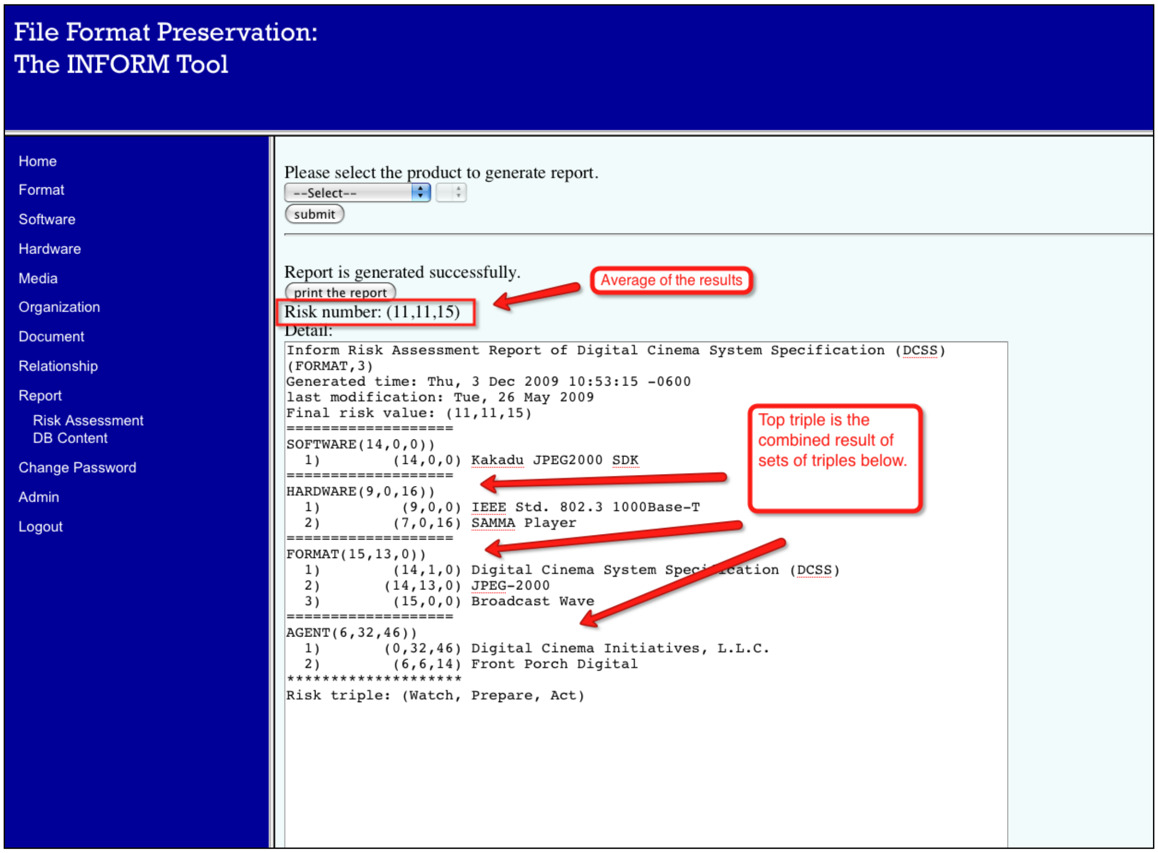
Risk	Probability of Loss	Size of Loss (Impact)
Royalties/license fees may be requested.	Very Low	Minor
Specification not available for independent inspection.	Very Low	Minor
Prior versions of the specifications are incompatible with each other.		Minor
Specification is very complex, large, ambiguous or poorly documented.		Minor
Specification is not widely accepted, either de jure or de facto.	Very Low	Minor
Specification is unique in its class and cannot be mapped onto another, or embedded metadata cannot be mapped onto other formats.	Very Low	Minor
Specification uses DRM schemes, signed envelopes, encrypted sections or watermarks.	Very Low	Minor
Staff with necessary expertise is not available.	Very Low	Minor
Testing procedures and staff are overwhelmed by rapidly changing specifications.	Very Low	Minor
Format specifications.	Very Low	Minor
Library specifications.	Very Low	Minor

Figure 12 – Format page in the INFORM tool. In addition, there are similar pages for Software, Hardware, Media, Organization, Document, and Relationship.



Probability of Loss		Size of Loss	
Very Low	below 1%	Minor	Insignificant Data Loss
Low	between 1 - 5%	Significant	Preventable Partial Data Loss
Moderate	between 6 - 10%	Serious	Preventable Complete Data Loss
High	between 11 - 25%	Very Serious	Unavoidable Partial Data Loss
Very High	Above 26%	Catastrophic	Unavoidable Complete Data Loss

Figure 13 – Tables defining levels of probability of loss and levels of size of loss. The left side of each table is what the user accesses in the pull-down menus.



Please select the product to generate report.

Report is generated successfully.

Risk number: (11,11,15)

Detail:

```
Inform Risk Assessment Report of Digital Cinema System Specification (DCSS)
(FORMAT,3)
Generated time: Thu, 3 Dec 2009 10:53:15 -0600
last modification: Tue, 26 May 2009
Final risk value: (11,11,15)
====================
SOFTWARE(14,0,0))
  1)        (14,0,0) Kakadu JPEG2000 SDK
====================
HARDWARE(9,0,16))
  1)         (9,0,0) IEEE Std. 802.3 1000Base-T
  2)        (7,0,16) SAMMA Player
====================
FORMAT(15,13,0))
  1)        (14,1,0) Digital Cinema System Specification (DCSS)
  2)       (14,13,0) JPEG-2000
  3)        (15,0,0) Broadcast Wave
====================
AGENT(6,32,46))
  1)       (0,32,46) Digital Cinema Initiatives, L.L.C.
  2)        (6,6,14) Front Porch Digital
********************
Risk triple: (Watch, Prepare, Act)
```

Average of the results

Top triple is the combined result of sets of triples below.

Figure 14 – Report generated by the INFORM tool drawn from data input by the user concerning a format.

6.7 Phase 2 - Research Protocol Development

The research protocol development phase occurred in parallel with software development. A key task was to contact personnel with digital preservation responsibilities at libraries and other organizations. Given the focus on audio and video formats, there were a limited number of potential contacts among the current NDIIPP partners. Eventually, we secured commitment to test the software from the following: Stanford University, UIUC, Florida Center for Library Automation (FCLA), HathiTrust, and the Library of Congress. Proceeding with initial research based on this small sample, our plan was also to query these contacts for additional ones and then snowball from that after initial data collection. However, while we anticipated some lag-time in response (because of the time-intensive nature of the INFORM tool), we did not expect the extent of delays that arose during this stage of the project. The busy schedules of the test users we engaged, along with the intricacies of learning the tool and working with it, were not conducive to enlarging our sample. Another key task was to secure IRB approval/exemption for the data-gathering process; below is a description of our approach, submitted to the IRB. Exemption was granted without reservation.

For our qualitative research approach we proposed a three-phase data collection process. In the first phase, once survey contacts were identified, and they agreed to participate, we planned to train participants on the survey tool. We viewed these experiences of initial training as data collection opportunities. We called these training sessions "walk-throughs," and to facilitate them we relied on WebHuddle[52], a virtual communication application for capturing user interactions with software as well as discussions with trainees. In preparing for this phase of research, we plotted out a walk-through of the tool, in order to highlight the functionalities of the tool clearly and accurately for the test subjects. Since use of the tool is time-intensive, we recommended to our test subjects that, after the walk-through of the tool, they then use it at their institutions to assess two to three data formats, at most, over the course of a couple of weeks—after which they would provide us with feedback on using the tool.

Feedback and follow-up characterize the second and third phases of our data collection. User feedback in essence made up the second phase, in which participants were observed on how they used the system we developed—how they applied it, for example, to assess similar and dissimilar data formats. In the third phase we did follow-up interviews (semi-structured, approximately two hours in length) with each institution (i.e., representative participant) engaged in our study. Our intent here was to elicit the sort of context informing risk assessment of data formats, with particular attention to the nature of the collection; the decision-making processes of the institution doing risk assessment; the preservation goals of the institution; and the sources of information available to them that allows them to perform a risk assessment. In Section 6.7.2 below we report on both these phases, since it was through semi-structured interviews that we obtained feedback. Once these phases of data collection were completed, a lengthy process of data analysis followed.

6.8 Phase 3 – Data Gathering and Analysis

6.8.1 Walk-throughs of the tool

During this phase of the project, we enlisted the participation of librarians and information professionals who have expertise in media preservation and digital preservation. We presented to each an overview, known as a walk-through, of the INFORM tool and its functionalities. These sessions were intended to give our test users an opportunity to get acquainted with the work we are doing and ask

[52] https://www.webhuddle.com/

questions about the tool and our project—essentially setting a context for their own individual interactions with the tool later.

The walk-throughs revealed some interesting questions, as well as pointed us to ways of improving the usability of the tool. Common questions that were posed touched on the following: the calculation of risk exposure (which involves using a MAX function); whether the tool comes pre-loaded with data about formats, to apply as guiding examples when using the tool; and whether our project is interested in local/institutional views of data format risk, or views of it in general. These questions led us to fine-tune our walk-throughs for future sessions. Based on these sessions, for example, we loaded the tool with a few examples, to give test users an idea of the kind of information being sought. In addition, all three participants expressed concern regarding the amount of time it would take to input data about preservation risks for two to three formats (our suggested number of formats), and a couple of participants cautioned to us that they might *not* know enough about digital formats. As a result, in the next round of user testing, we engaged participants who are more established in the digital preservation realm and thus are more experienced.

6.8.2 Feedback on the tool

In our follow-up interviews with test users we were interested in finding out—among other things—how well the tool works, how it could be improved, and whether the user's institution or organization would actually use the tool (either in its current, or improved, state). In general, our users found that the tool works well, but they would have liked visual aids such as diagrams and user interface aids, such as pop-up windows providing explanatory text about particular menu items. Participants also said they would have benefited from a manual or some kind of user documentation about the tool. Another common observation was the expectation that it would be possible to see records input by previous users and thus view the INFORM tool as a knowledge base. Our test users believed a tool like this would be useful at their institutions, but they harbored concerns about the length of time it took them to input information about a format and about the number and variety of formats – i.e., sometimes, there are not many formats that an institution works with, nor do the formats vary widely.

6.8.3 Analysis and suggestions for future work

While testers of the INFORM tool found the basic user interface and workflow of the tool relatively easy to navigate, they had significant problems with various different aspects of the INFORM methodology itself and obtaining and creating the information necessary to complete an assessment for a given file format. One of our tester's comments provided a remarkably apt summary of many of the testers' reaction to the INFORM tool: "It does a good job guiding the user through the process, but it's pretty subjective."

Many of the testers found various data elements requested or required by the INFORM tool difficult or impossible to provide. For the various descriptive metadata components providing information about a format, this is not a significant problem; information regarding a file format such as the date it was withdrawn from circulation could be provided from a central repository such as the UDFR. The actual assessment information, unfortunately, also proved problematic for the majority of our testers:

- "I was guessing on some of these things." [regarding file format assessment information]
- "It's hard to hazard a guess as to what the impact would be. You could imagine it would be difficult…. one thing that struck me in so many of these screens -- in so many cases the answer is I don't know and I don't know when I will know, and it may take too long to research it."
- "What does probability of loss mean? Is that immediate loss? Loss down the road? It's hard to mentally wrangle because it's hard to know what it's referring to."
- "I would have a hard time doing a quick and accurate estimate of these things. I would need to spend a fair amount of time talking to the developers of the software to find out what its dependencies were and how it operated."
- "There's something a little artificial about assigning probabilities that gives you a false precision that isn't that useful."

Somewhat more disturbing from the point of view of trying to establish a common shared metric for risk assessment was the fact that many of the individuals evaluating the INFORM tool commented on the fact that a risk evaluation they might conduct at their own institution would be based on a number of local contextualizing factors that a numeric metric would obscure. Some institutions also indicated that the ways in which they evaluated probability of loss might differ from how other institutions evaluated probability of loss.

- "There's a real tension between evaluating for your own institution and evaluating for part of a Delphi approach to assessing risk for everyone."
- "There's nothing here that talks about the amount of data I have in a given format. This is designed for a different metric, but it's hard to shift gears and say that the number of files we have doesn't matter." [in regards to assessing impact]
- "Is there anything in here about the content? If the content is a rare J. D. Salinger interview, my acceptable risk level changes."
- "There's different perspectives on risk. The only situation in which we can control what file formats come in the front door is ETDs, because the universities have leverage. So, we just put out a risk assessment that details what the risks are for different formats for ETDs. That risk assessment is

different than the risk assessment we're doing here [with the INFORM tool], because we know we can normalize it."

- "Migratability [sic] is what we're talking about. There's the risks of the format, and there's the risk of losing your data. When you have the ability to migrate, the format may be risk, but the data isn't at risk."

Some of the users also were concerned about the time involved in conducting risk assessments and wondered whether this approach was scalable.

- I'm not sure how scalable this is. In a large institution like [ours], you could actually have dozens or hundreds of digital formats. Given the time consuming nature of risk assessment, it seems like the time commitment to doing all the assessment work would be difficult."
- "Right now what we're currently managing in our preservation system, the number of formats is not that wide. It's a dozen, maybe two dozen. Of course we know that that's all going to change as our services open up to more and more people on campus, and we start trying to address more heterogeneous collections. But at this point, the number of formats that we're watching is very low. I would think that this tool would , as the number of formats in your care started increasing, the role of this tool would become more important. It greatly facilitates your ability to manage those in a structured way. But of course as the number of formats proliferate, just keeping the information in a tool like this up to date would become more challenging from a resource point of view."

Our testers also found the actual evaluations output by the INFORM methodology difficult to interpret. This is particularly worrisome given that they were responsible for the input of the data upon which the evaluation was based.

- "Understanding the triad results is difficult. I reviewed the methodology to understand the method behind it, but on its own a report doesn't really help that much. It might be better if you could compare those numbers against other results, and seeing how differences in your assessment of risk factors influence the outcome would be helpful from an educational point of view…. I found the results difficult to interpret so it was difficult to decide whether they match my existing sense."
- "We kind of fell apart at the report stage. The risk exposure result still mystifies us. If you want compare this format to another one, it's OK, but if I want an assessment that doesn't involve looking at other formats, this doesn't work."
- "The numbers that come out of the reports are somewhat opaque."

Interestingly, while the testers had doubts about the value of the results provided by the INFORM tool, some of them commented favorably on the process of using it. By providing a structured means of examining the various factors involved in evaluating a file format, INFORM did provide assistance in conducting an examination of file format's risks, even if the output format was less than desirable.

- "It was really useful for me and to us when we sat down and thought through all of the dependencies. It's really useful to map it out…. The web of requirements and interdependencies that exists between operating systems, and software and open source software continues to amaze me."

Taken as a whole, the testers' comments throw doubt upon the usefulness of the INFORM methodology as originally proposed. Risk assessments of file formats will be made within the context of a local institution's available resources and existing practices, and a quantitative metric of a format's risk will inevitably hide the local contextualizing factors influencing a given institution's assessment. Sharing of such assessments could actually be counterproductive to the digital preservation community as a whole. It is easy to imagine a situation in which the time and resources necessary to conduct such detailed risk assessment result in only larger, more well-funded agencies producing such assessments. Those institutions' assessment of risk, based upon their own institutional context, could result in risk assessments that differ significantly from those that might be produced by institutions with fewer resource to bring to bear. But the lack of context surrounding those assessments within INFORM could make the assessments in a shared resource appear authoritative for all. INFORM simply doesn't provide enough information on how assessments are conducted to allow one party to evaluate another's assessment and its meaning for their own institution.

Based on our research, we cannot recommend the INFORM methodology as originally proposed as a tool for the digital preservation community. However, based upon the comments from our evaluators, there are some steps that could be taken that we believe would be of value to digital preservationists. While the quantitative metrics of INFORM are not necessarily that valuable even for one institution and have real problems in terms of providing shared data, the larger evaluative framework proposed in INFORM, particularly the interdependencies between file formats, software, hardware, media and organizations, was commented upon by several of our evaluators as a useful way to think about risk issues with respect to file formats. It would be valuable if file format registries such as the UDFR could include within the information they record details about known interdependencies. The Digital Cinema Specification, for example, relies upon the JPEG2000 still image format for individual frames within a movie. Having such dependencies recorded in the UDFR and available for consulting by the digital preservation community would obviously be of value. Some of the evaluators also commented that while the INFORM tool might not provide the information they want regarding risk assessment of a format, they are very interested in the larger community's opinion regarding file formats' risks, but with the qualifier that they want to know why a particular institution considers a format risky (or not). Tools for format registries which allow for the input of more qualitative assessment data from different preservation repositories we believe would be seen as very valuable.

7. References

Carlson, A. J.; Cumby, C. M.; Rizzolo, N. D.; Rosen, J. L.; & Roth, D. (2004). *SNoW user manual*. Urbana, IL: University of Illinois Department of Computer Science, Cognitive Computation Group. Available at http://l2r.cs.uiuc.edu/~cogcomp/software/snow-userguide/userguide.html.

Cheney, J.; Lagoze, C.; &Botticelli, P. (2001). *Towards a theory of information preservation.* Technical Report TR2001-1841, Cornell University, Ithaca, NY, 2001.

Cruse, P.; & Sandore, B., eds. (2009). Library Trends 57 (3) Winter 2009: Library of Congress NDIIP Program. Available at http://www.ideals.illinois.edu/handle/2142/13586

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia. In *EMNLP 2007: The Joint meeting of the Conference on Empirical Methods in Natural Language Processing*.

Devlin, K. (1995). *Logic and Information*. Cambridge University Press.

Downing, P. (1977). On the creation and use of English compound nouns. *Language* 53:810-842.

Dresher, M. (1981). *The mathematics of games of strategy.* Dover Publications.

Dubin, D.; & and Birnbaum, D. J. (2008). Reconsidering conventional markup for knowledge representation. Presented at Balisage: the Markup Conference, Montreal., August 2008.

Dubin, D. (2005). Unpacking the interpretation of METS markup. Presented at the Digital Library Federation Fall Forum, Charlottesville, VA, November 2005.

Dubin, D. (2007). Instance or expression? another look at reification. Presented at Extreme Markup Languages 2007, Montreal., August 2007.

Dubin, D.; Futrelle, J.; & Plutchak, J. (2006). Metadata enrichment for digital preservation. In: *Extreme Markup Languages 2006, Montréal, Québec, August 7-11, 2006*. Available at http://hdl.handle.net/2142/9461.

Dubin, D.; Futrelle, J.; Plutchak, J.; & Eke, J. (2009). Preserving meaning, not just objects: semantics and digital preservation. *Library Trends*, 57(3):595--610, 2009.

Gabrilovich, E. & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research, 34(1)*. 443-498.

Gardner, H.; Kleiner, F. S.; & Mamiya, C. J. (2005). *Gardner's art through the ages*. Belmont, CA: Thomson/Wadsworth.

Godby, C. J.; Hswe, P.; Jackson, L.; Klavans, J.; Ratinov, L.; Roth, D.; & Cho, H. (2009). Who's who in your digital collection: developing a tool for name disambiguation and identity resolution. In *Proceedings of the 2009 Chicago Colloquium on Digital Humanities and Computer Science*. Publication pending. Available at http://hdl.handle.net/2142/15393.

Grishman, R. & Sundheim, B. (1995). Design of the MUC-6 evaluation. In *Proceedings of the 6th Conference on Message Understanding. Columbia, Maryland, November 06 - 08, 1995*, 1-11. Morristown, NJ: Association for Computational Linguistics.

Guarino, N.; & Welty, C. A. (2000). A formal ontology of properties. In *EKAW '00: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management, London, UK, 2000*, 97-112. Springer-Verlag.

Habing, T.; Eke, J.; Cordial, M.A.; Ingram, W.; Manaster, R. (2009). Developments in digital preservation at the University of Illinois: the hub and spoke architecture for supporting repository interoperability and emerging preservation standards. *Library Trends* 57(3), Winter2009, 556-579.

Jackson, L. S. (2003). Preserving state government web publications -- first-year experiences. In: *National Conference on Digital Government Research (DGO-2003), Boston, MA, May 18-21, 2003*, Digital Government Research Center, 109-114. Available at http://hdl.handle.net/2142/16400.

Jackson, L. S. (2005). Difficulties in electronic publication archival processing for state governments. In: *1st International Conference on Universal Digital Library, ICUDL 2005*, 175-185. Available at http://hdl.handle.net/2142/16401.

Jackson, L. S. (2010a). *Testing the extracting metadata for preservation project's named entity recognizer on metadata.* Technical Report #ISRN UIUCLIS--2010/1+EAP. Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Available at http://hdl.handle.net/2142/15401.

Jackson, L. S. (2010b). *Functional genre in Illinois State Government digital documents.* Technical report #ISRN UIUCLIS--2010/3+EAP. Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Available at http://hdl.handle.net/2142/15475.

Klavans, J.L.; Abels, E.; Lin, J.; Passonneau, R.; Sheffield, C.; & Soergel, D. (2009). Mining texts for image terms: the CliMB project. In *Digital Humanities 2009. Conference Abstracts. University of Maryland, College Park, USA. June 22 – 25, 2009*, 184-186. College Park, MD: Maryland Institute for Technology in the Humanities (MITH). Available at http://www.umiacs.umd.edu/~jimmylin/publications/Klavans_etal_DH2009.pdf.

Klavans, J. L. & Resnik, P. (eds.) (1996). *The balancing act: combining symbolic and statistical approaches to language.* Cambridge, MA: MIT Press.

Kripke, S. (2000). *Naming and necessity*. Oxford: Blackwell.

Lafferty, J.; McCallum, A.; & Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, Brodley, C. E. & Danyluk., A. P. (Eds.). San Francisco: Morgan Kaufmann.

Lamarque, P. (2002). Work and object. *Proceedings of the Aristotelian Society*, 102(1):141--162.

Levinson, J. (1990). Music, art, and metaphysics: essays in philosophical aesthetics, chapter 4, 63-88. Ithaca, NY: Cornell University Press.

Li, X. & Roth, D. (2005). Discriminative training of clustering functions: theory and experiments with entity identification. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL). 29-30 June 2005. University of Michigan, Ann Arbor, Michigan, USA.* Madison, WI: Omnipress, Inc.

Mihalcea, R. & Csomai, A. (2007). Wikify! Linking documents to encyclopedic knowledge. In *CIKM 2007: ACM Sixteenth Conference on Information and Knowledge Management.*

Moreau, L.; Plale, B.; Miles, S.; Goble, C.; Missier, P.; Barga, R.; Simmhan, Y.; Futrelle, J.; McGrath, R. E.; Myers, J.; Paulson, P.; Bowers, S.; Ludaescher, B.; Kwasnikowska, N.; Van den Bussche, J.; Ellkvist, T.; Freire, J.; & Groth, P. (2008). *The open provenance model (v1.01)*. Technical report, University of Southampton, July 2008.

Renear, A. H.; & Dubin, D. (2007). Three of the four FRBR group 1 entity types are roles, not types. In Andrew Grove (ed.), *Proceedings of the 70th Annual Meeting of the American Society for Information Science and Technology, Medford, NJ, 2007*. Information Today, Inc. Available at http://hdl.handle.net/2142/9094.

Renear, A.; Dubin, D.; & Wickett, K. (2008). When digital objects change - exactly what changes? Presented at the 2008 Annual Meeting of the American Society for Information Science and Technology, Columbus, OH, October 2008.

Sang, E. F. T. K. & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: language independent named entity recognition. In *Proceedings of Seventh Conference on Natural Language Learning (CoNLL). 31 May-1 June 2003. Edmonton, Canada*, 142-147. Morristown, NJ: Association for Computing Machinery.

Searle, J. R. (1999). *Mind, language, and society: philosophy in the real world*. New York: Basic Books.

Smith, B. (2007). Searle and de Soto: the new ontology of the social world. In Barry Smith, David M. Mark, and Isaac Ehrlich, editors, *The Mystery of Capital and the Construction of Social Reality*, 35--51. Chicago/La Salle IL: Open Court.

Smith, B.; & Searle, J. (2003). The construction of social reality: an exchange. *American Journal of Economics and Sociology*, 62(1):285--309.

Stanescu, A. (2004). Assessing the durability of formats in a digital preservation environment. *D-Lib Magazine, 10*(11). Retrieved November 30, 2009, from http://www.dlib.org/dlib/november04/stanescu/11stanescu.html.

Ratinov, L. & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *CoNLL-2009: Thirteenth Conference on Computational Natural Language Learning*, 147-155. Boulder, CO: Association for Computational Linguistics. Available at http://aclweb.org/anthology-new/W/W09/W09-1119.pdf.

Renear, A. H.; Wickett, K. M.; Urban, R. J.; Dubin, D.; & Shreeves, S. L. (2008). Collection/item metadata relationships. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*. Available at http://dcpapers.dublincore.org/ojs/pubs/article/viewPDFInterstitial/921/917.

Toms, E. G. (2001). Recognizing digital genre. *Bulletin of ASIS&T*, 27(2).

Van de Sompel, H. & Beit-Marie, O. (2001). Open linking in the scholarly information environment using the OpenURL framework. *D-Lib Magazine*, 7 (3). Available at http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html.

Appendix A: EMP Project Annotation Instructions

CoNLL Categories:

Locations (LOC):

- roads (streets, motorways)
- regions (villages, towns, cities, provinces, countries, continents, dioceses, parishes)
- structures (bridges, ports, dams)
- natural locations (mountains, mountain ranges, woods, rivers, wells, fields, valleys, gardens, nature reserves, allotments, beaches, national parks)
- public places (squares, opera houses, museums, schools, markets, airports, stations, swimming pools, hospitals, sports facilities, youth centers, parks, town halls, theaters, cinemas, galleries, camping grounds, NASA launch pads, club houses, universities, libraries, churches, medical centers, parking lots, playgrounds, cemeteries)
- commercial places (chemists, pubs, restaurants, depots, hostels, hotels, industrial parks, nightclubs, music venues)
- assorted buildings (houses, monasteries, nurseries, mills, army barracks, castles, retirement homes, towers, halls, rooms, vicarages, courtyards)
- abstract ``places'' ("the free world")

Miscellaneous (MISC):

- words of which one part is a location, organization, miscellaneous, or person (e.g. Indonesian martial arts)
- adjectives and other words derived from a word which is location, organization, miscellaneous, or person (e.g. Socratic method and Hippocratic oath)
- religions
- political ideologies
- nationalities
- languages
- programs
- events (conferences, festivals, sports competitions, forums, parties, concerts)
- wars
- sports related names (league tables, leagues, cups)
- titles (books, songs, films, stories, albums, musicals, TV programs)
- slogans
- eras in time
- years (e.g. 2001)

- types (not brands) of objects (car types, planes, motorbikes)

Organizations (ORG):

- brands
- companies (press agencies, studios, banks, stock markets, manufacturers, cooperatives)
- subdivisions of companies (newsrooms)
- political movements (political parties, terrorist organizations, governments, government-sponsored actions)
- government bodies (ministries, councils, courts, political unions of countries (e.g. the [U.N.]))
- publications (magazines, newspapers, journals)
- musical companies (bands, choirs, opera companies, orchestras)
- public organizations (schools, universities, charities)
- other collections of people (sports clubs, sports teams, associations, theater companies, religious orders, youth organizations)

Persons (PER):

- first, middle and last names of people, animals and fictional characters
- aliases , titled names in common use

Markup Convention:

[LOC/ORG/PERS/MISC Named Entity [explanation]]

- Insert rationale for the marked words in inside brackets
- Named entities do not overlap.
    - e.g., [MISC The Oprah Winfrey Show [program]] NOT person
- Examples:
    - ORG
        - [ORG Coca-Cola []]
        - The [ORG National Organization for Women [nonprofit]]
        - [ORG Department of Labor [govt body]]
    - PER
        - [PER President Obama [alias/proper name]] called on citizens to
        - [PER George Eliot [alias]]
        - [PER Superman [fictional name]]
        - [PER Mauricio Hernandez [full name]]’s family reunion
    - LOC
        - [LOC Jura [mountain range]]
        - [LOC Swiss [country]] cows roam the streets

- The accident occurred on the [LOC A4 [motorway]]
- [LOC Japan [country]]’s coins feature
  - BUT [MISC Thailand’s Royal Family [adjective in the name + location]]
- MISC
  - [MISC Hindu [religion]]
  - [MISC Super Bowl [event]]
  - [MISC The Oprah Winfrey Show [TV program]]
  - During the [MISC Great War [war]]

Appendix B: OCLC Repository of EMP Project Source Code, Training Data, Testing Data, and Result Sets

Storage facilities for EMP project source code, training data ("gold standard"), testing data, and result sets have been constructed at OCLC using Subversion version control software[53] and subsequently hosted on the Web using Google Project Hosting.[54] Modified NER tool source code, convenience scripts, training data, testing data, and results are available for download at this facility.

The NER tool software as originally created at UIUC also continues to be available for download, along with many companion works, via the Cognitive Computation Group website[55] at the University of Illinois Department of Computer Science.

Both the OCLC software and the UIUC software are planned to undergo further development beyond the NDIIPP-2 research grant.

[53] http://subversion.tigris.org/

[54] http://code.google.com/p/emp-suite/

[55] http://l2r.cs.uiuc.edu/~cogcomp/