Summary

The exponential growth of information, especially with the evolution of Internet, poses many challenges which librarians and information professionals have been grappling with for decades. Traditional method of providing access to “right information to the right user at the right time” through use of catalogues has given way to online search engines.

Even though online search engines provide ease and speed in resource discovery they lack precision and increase noise. For example, a search in Google Scholar with the keyword ‘India’ brings up the first result as a citation of *Ultra-violet and visible spectroscopy: chemical applications* whereas in Scirus, the first three results are office location pages of some companies. The same search in OCLC Worldcat, which is a metadata based search engine, shows first result as a book on Indian history.

If a metadata based service can be built that integrates multiple data formats, multiple protocols, and multiple metadata standards it would be a boon to libraries and the research community. Such an effort will need bringing non-structured metadata sources into structured formats. This study is an attempt to identify the challenges and demonstrate the possibilities of achieving such objective.

Saiful Amin
Research Student

Dr. A. R. D. Prasad
Research Guide
Users do not have the time to go through every possible source of information to find out relevant items for their research. The normal behaviour of any user is to go to his/her favourite resource and collect as much information as possible. It not only limits their scope of search but also renders various important resources, whether free or commercial, unused. To facilitate better and more relevant online resources to users there is a need for an integrated interface to all possible online repositories in any subject area.

Any retrieval system is as good as the metadata used for describing the content. The web crawlers and search engines tend to index only the full-text of online content. This contributes to the noise in the search results, especially, while looking for scholarly information.

1. Scope of Work

The present work is in the area of resource discovery through metadata harvesting. The concept of metadata harvesting is a recent concept which deals with aggregation of metadata from various online resources, bibliographic and full-text, as an efficient way to provide federated searching across archives. This concept was promoted by Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

The present work enables creation of models for creating metadata from repositories on-the-fly even when the publisher does not follow any existing metadata standard. The study uses HTML heuristics for creating metadata from existing descriptions of the work available on the publisher websites.
2. Objectives

The objectives of the present work were the following:

- Develop techniques for harvesting and extracting metadata from diverse information resources.
- Define specific types of information resources and the methods to aggregate their metadata.
- Facilitate creation of subject specific metadata repositories for all types of online scholarly information resources.
- Automate article-level indexing on a regular basis.

3. Hypotheses

The present work was based on the following hypotheses:

- Scholarly publishing has mostly been done on legacy and proprietary systems without much respect to any metadata standard.
- Metadata-based retrieval system eliminates noise as well as provide better and relevant search results.
- Heuristics-based system can be used effectively to create metadata from resources that do not follow any existing metadata standard.

4. Methodology

The first stage of the research involved studying existing systems that index multiple repositories for scholarly publication, such as, OAIster, Scirus.
Google Scholar, etc. The outcome of this stage was the identification of certain limitations of such services.

The second stage of this work was to identify various metadata structures used for indexing e-journals articles and other scholarly information resources. In this stage, the use of Dublin Core, MARCXML, RSS and other XML-based standards was explored to find the most suitable format for representation of article-level metadata. It was decided to use Dublin Core format in XML since it serves as a de facto standard.

The third stage was to develop techniques to harvest metadata from disparate online repositories and individual publishers. This stage explored various innovative techniques for extraction of metadata not only from metadata archives but also creating metadata on-the-fly from full-text resources available in the public domain. The challenge was to extract metadata from resources which do not follow any set structure in describing their content. Crosswalk tools were developed to map various existing metadata formats into the preferred Dublin Core format.

As a proof of concept of the project a prototype was developed using the techniques identified during the project.

5. Chapter Organization

This study is organized into the following chapters:

**Chapter 1: Introduction**

This introduction chapter covers an overview of the entire research work, including a brief introduction, need for the present research, the scope, objectives, hypotheses and methodology used for the research work.

**Chapter 2: State of the Art**

This chapter is a state of the art report on the current state of discovery
services of scholarly resources, with emphasis on metadata support on those services.

Chapter 3: Metadata Sources
This chapter provides description of content sources from which metadata can be collected or extracted.

Chapter 4: Metadata Collection
This chapter provides strategies regarding the metadata collection from disparate resources. It focusses on mapping of various metadata formats from multiple type of resources using multiple protocols.

Chapter 5: Design and Demonstration
This chapter describes the development of the prototype using the strategies as detailed in previous chapters. It also includes the demonstration of the prototype.

Chapter 6: Conclusion
This concluding chapter presents the observations and suggestions made under various chapters.