The goal of this research is to develop a tri-dimensional metadata model and implement this model through the ScholarWiki system to combine the machine-induced, user-enhanced metadata for more effective knowledge discovery and information retrieval. The tri-dimensional model captures the Structural, Descriptive, and Referential (SDR) metadata and incorporates them into a social media platform—ScholarWiki system. By allowing low-barrier participation, scholars (both as authors and users) can participate in the knowledge and metadata editing and enhancing process and benefit from more accurate and effective information retrieval. The ScholarWiki system utilizes machine-learning techniques that can automatically produce self-enhanced metadata through learning the structural metadata that scholars contribute. The cumulated machine learning will add intelligence to automatically enhance and update the publication metadata Wiki pages.

Keywords
ScholarWiki, Metadata, Information Retrieval, Human Intelligence, Human Computing

INTRODUCTION
The ability of discovering, integrating, and re-using relevant scholarly outputs from prior studies is critical for innovative research (Shotton et al 2009). Such ability is largely facilitated by the metadata that represents scholarly outputs. Traditionally, metadata has been primarily focused on descriptive representation with elements such as title, author, publishing, and subject keywords. Realizing that descriptive-centric metadata has become increasingly inadequate as the complexity and volume of scholarly output grows, database vendors and information systems of scholarly output developed innovative mechanisms to address these challenges. For example, most indexing databases such as ACM Digital Library (http://portal.acm.org/) and SciVerse Scopus (http://www.info.sciverse.com/scopus) allow for reference linking. Meanwhile, some more recent researches begin to be aware of the importance of the structural metadata, such as structural abstract. These new developments enhanced the retrieval and use of scholarly output through providing referential and structural metadata, but they are scattered in different systems and products and become static information once the metadata is generated. By static we mean that the scholars have no way to modify or edit the information about their publications once they are handed over to the publisher. While descriptive, referential, and structural metadata can enhance the discovery and use of scholarly output, current systems allows little or no scholar’s involvement in creating, modifying, and editing metadata for their publications. This paper proposes an ontology model for integrating the Structural, Descriptive, and Referential (SDR) metadata into one platform—Wiki—to allow for scholar’s participation in creating and editing SDR metadata for their publications. Figure 1 shows this.

Fig 1. Metadata creation and enhancement process with participation by user, machine and expert
metadata creation and enhancement system’s architecture, in which domain experts, users, and a machine learning model work together to create and enhance metadata for scholarly outputs. The ScholarWiki system functions as the interface for users and authors to actively participate in the metadata creation, editing, and maintenance processes.

REVIEW OF RELEVANT RESEARCH

Metadata describes characteristics of resources for discovery and organization purposes and is often defined as “data about data”. Its importance in managing and using digital resources can never be overemphasized, especially in the cyberinfrastructure-enabled, distributed collaboration between scientists, a.k.a., “e-Science” (Goble 2005). As massive, diverse research outputs complicate information management and discovery in the highly distributed, computing-intensive e-Science environment, the generation and maintenance of metadata needs innovative approaches to keep it updated, enriched, and interlinked. In order to accomplish this, we need to have a thorough understanding of what has been done and what are some of the limitations of existing approaches.

Most metadata development has inherited the description tradition from the library community. Typical descriptive data fields include title, source, date, publisher, keyword, and abstract as specified in Dublin Core Metadata Element Set¹ and many other standards. Although metadata for scholarly research outputs also follows this tradition, it has become clear that the changing research paradigm, namely e-Science (Hey, Tansley, & Tolle, 2009), demands more effective models and tools to support the data-intensive, distributed science research collaboration.

Citations in research papers represent a non-descriptive tradition in metadata developments. The Citation Typing Ontology (CITO) (Shotton, 2008) is an example. It employs an ontology-based metadata model in RDF format to describe the intent of citations, e.g., confirm, correct, credit, critique, and disagree with. In addition, some description models have been given new life by structuring the properties in an ontological language, e.g., the Bibliographic Ontology (BIBO) (Dabrowski, Synak, & Kruk, 2009) and the Scholarly Works Application Profile (SWAP).

One important dimension of scholarly publications that has been largely left out of the descriptive tradition is the structural metadata. By structural we mean the rhetoric structure in scholarly publications that signify research problems, theories/models, hypothesis, methodologies, findings, and conclusions. Many biomedical research articles contain structured abstracts to facilitate search and browsing, e.g., The New England Journal of Medicine (http://www.nejm.org). Structured metadata can make it easier for harvesting the vast metaknowledge across

¹ http://purl.org/dc

THE SDR MODEL

The SDR metadata model builds on three dimensions — structural, descriptive, and referential. Two ontologies are developed to model the three dimensions: the Structural Ontology at domain level and the Descriptive & Referential Ontology at publication level (Fig.2).

The domain ontology characterizes the structural elements most important for representing the common publication features in a specific domain. For example, the domain “Information Retrieval” has four essential structural elements: “Research Question”, “Methodology”, “Dataset”, and “Evaluation”. While some structural elements are common across domains, others are unique in their own field. The publication ontology describes a publication in its domain from descriptive and referential perspectives. Unlike existing metadata schemas that contain descriptive information only, this model has embedded structural information (as the fields in Fig 2 with *) in the publication ontology. Take the domain “Information Retrieval” as an example again: a keyword for a publication can be described as the relationship shown in Fig. 3, “Keyword 1” is given the context of “Methodology”. Similarly, we can also easily define a citation as about the “Dataset” or “Evaluation” of this publication.

The SDR model distinguishes between referential metadata and reference metadata. Today’s scholarly outputs tend to be not only interdisciplinary but also diverse in types and formats. Conference proceedings may be accompanied by presentation slides and/or podcasts, and a journal article may have supplemental materials such as datasets, tables, and charts, which may or may not be stored at the same network location. We consider these related and supplemental materials as part of the referential information of scholarly outputs. The SDR model will help capture the referential metadata by using the “resource” element. Since the information about referential materials tends to be
dynamic and controlled by scholars, our participatory ScholarWiki would provide a tool for referential metadata generation.

Because of space limitation, we cannot present the details of the SDR ontology model. Please find more details at http://discern.uits.iu.edu:8826/webprotege/#ScholarWiki

THE SCHOLARWIKI SYSTEM
Implementing the SDR model on a wiki platform is necessary for testing the model’s feasibility for a collaborative, distributed knowledge and information representation and enhancement system. Research has discovered that, while some metadata elements such as title and creator are popular and frequently used as access points, others are rarely used (Godby, 2004) due to the difficulty in identifying the values for these elements and the unknown or little return on time invested in creating them. Developing SDR ontology for each research domain is a daunting task and the cost of hiring domain experts to accomplish this would be prohibitively high. Given that the ontologies are dynamic and evolve over time, it would be ideal to have scholars and librarians modify or update the metadata any time after it is created. If a publication introduces an innovative algorithm and one year later the open source version is released to implement this algorithm, the researcher should have the access to update the referential metadata. The ScholarWiki system will be the platform for bi-directional metadata communication (from system to authors and users and vice versa).

The metadata creation and enhancement process in ScholarWiki can be generated in several stages. Initially at the expert stage, domain experts create domain ontology and provide sample structural keywords and other structural metadata as necessary to some random publications in this domain. In the second (system) stage, a machine-learning algorithm is deployed to learn from the domain expert’s coding results. The system will then generate Wiki template for the domain ontology, as Figure 4 shows. Meanwhile, a list of Wiki pages would be generated automatically for each publication that will bear the structural metadata in the form of a domain Wiki template. Fig. 5 shows the Wiki pages generated automatically by the inference algorithm (circled part is the structural metadata with Wiki template). Finally, at the user stage, users access and edit the Wiki pages to enhance or tweak the metadata presented on the pages. The machine-learning algorithm will learn the user’s editing behavior to improve its performance. The updated machine learning model will re-generate higher quality Wiki pages periodically and this process will repeat itself as operation routines. Another example page is available at: http://scholarwiki.indiana.edu/wiki/index.php?title=Using_searcher_simulations_to_redesign_a_polyrepresentative_implicit_feedback_interface

The ScholarWiki will automatically generate a list of Author Wiki and Keyword Wiki pages for quick access to resources. Each Author Wiki page presents the topics and exemplar publications that the author has published and the topics of these publications are automatically ranked by citation counts. Each Keyword Wiki page indexes the most notable authors, most cited publications and keywords related to these publications. Later in the next stage, the machine-learning algorithm will learn these patterns and features automatically and add the new updates to the author and keyword Wiki pages. For instance, in the information retrieval domain, machine learning can tell each author’s most salient and most cited research question, and the frequently used methodologies and datasets. Of course, all the metadata can be edited and enhanced by users through Wiki pages.

CONCLUSION
This research contributed to three fronts:

1. an innovative model that describes the structural, referential, and descriptive metadata of academic domain and publications, which forms the foundation of our system development;
2. a novel methodology to create and enhance metadata for scholarly outputs, which integrates the intelligence of domain expert, machine and user; and
3. a ScholarWiki system that implements the abovementioned methodology.
These three areas of work together build a healthy workflow to enable bi-directional metadata generation and enhancement. This novel approach promises an open, participatory process for organizing, disseminating, and sharing scholarly outputs.

We plan to continue to improve the ScholarWiki system by generating high quality Wiki pages. For instance, for each keyword Wiki page, a “mini review paper” for this keyword will be automatically generated. Some human readable paragraphs will be presented on the Wiki page along with some most influential paper links. Of course, every page in ScholarWiki could be enhanced by user, and machine will have the opportunity to learn user’s editing behavior.

REFERENCES