MESUR: usage-based metrics of scholarly impact.

Johan Bollen
Digital Library Research & Prototyping Team
Los Alamos National Laboratory
Los Alamos, NM 87545
jbollen@lanl.gov

Marko A. Rodriguez
Digital Library Research & Prototyping Team
Los Alamos National Laboratory
Los Alamos, NM 87545
marko@lanl.gov

Herbert Van de Sompel
Digital Library Research & Prototyping Team
Los Alamos National Laboratory
Los Alamos, NM 87545
herbertv@lanl.gov

ABSTRACT

The evaluation of scholarly communication items is now largely a matter of expert opinion or metrics derived from citation data. Both approaches can fail to take into account the myriad of factors that shape scholarly impact. Usage data has emerged as a promising complement to existing methods of assessment but the formal groundwork to reliably and validly apply usage-based metrics of scholarly impact is lacking. The Andrew W. Mellon Foundation funded MESUR project constitutes a systematic effort to define, validate and cross-validate a range of usage-based metrics of scholarly impact by creating a semantic model of the scholarly communication process. The constructed model will serve as the basis of a creating a large-scale semantic network that seamlessly relates citation, bibliographic and usage data from a variety of sources. A subsequent program that uses the established semantic network as a reference data set will determine the characteristics and semantics of a variety of usage-based metrics of scholarly impact. This paper outlines the architecture and methodology adopted by the MESUR project and its future direction.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks; H.2.8 [**Database Applications**]: Data mining; H.3.7 [**Digital Libraries**]: Collection

General Terms

Digital libraries, usage data, impact, scholarly evaluation, semantic networks, modeling, RDF, OWL, architecture, standards

1. INTRODUCTION

The quantitative assessment of scholarly communication items and the agents responsible for their production is a matter of increasing importance within a continuously expanding and diversifying scholarly landscape. It has now become common to use citation data to assess the impact of journals, journal articles, their authors, and the institutions they are affiliated with. Such assessments frequently influence funding and promotion decisions [7, 33, 36].

Scholarly communication is a multi-phased process that begins with the germination and development of an idea, resulting in its publication in a peer-reviewed journal and subsequent citations to that original publication. Citations therefore reflect the formal end-result of this process. However, throughout all phases of the scholarly process, scholarly communication items are downloaded, read, and otherwise consulted. It is therefore reasonable to expect that the analysis of usage data can lead to a different and potentially more complete perspective on scholarly impact than citation data alone can provide:

- Usage precedes citation, thereby serving as an earlier indicator of scholarly impact [13].
- Usage can be recorded for a range of scholarly communication items that extends beyond journal articles, and for a community not restricted to authors of journal articles.
- Usage can reflect many different aspects of scholarly impact by being recorded at multiple locations across the spectrum of digital information services, e.g. publisher services, institutional repositories, institutional link resolvers, etc.

Usage data thus holds the promise of greatly expanding the possibilities for scholarly assessment. For example, instead of evaluating scholars only by the number of citations their articles receive in the published literature, their work could in addition be evaluated in terms of how well-read their articles are among students and practitioners. The potential value of usage data is illustrated by numerous publications that find intriguing overlaps as well as discrepancies between citation- and usage-based scholarly impact [3, 6, 13, 25, 26]. However, usage data and corresponding metrics of scholarly impact have not yet made inroads as reliable and community-accepted components of the toolkit available for scholarly assessment.

The reasons for this lack of acceptance are manifold but the most important ones can be summarized as follows:

1. Obtaining usage data: privacy issues and the lack of usage recording standards as well as standards for the aggregation of usage data make it difficult to establish reliable, aggregated usage data sets.
2. Sampling: usage data is generally recorded by information services dedicated to a particular user community.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL '07 Vancouver, CA
Copyright 2007 ACM X-XXXXX-XX-X/XX/XX ...$5.00.

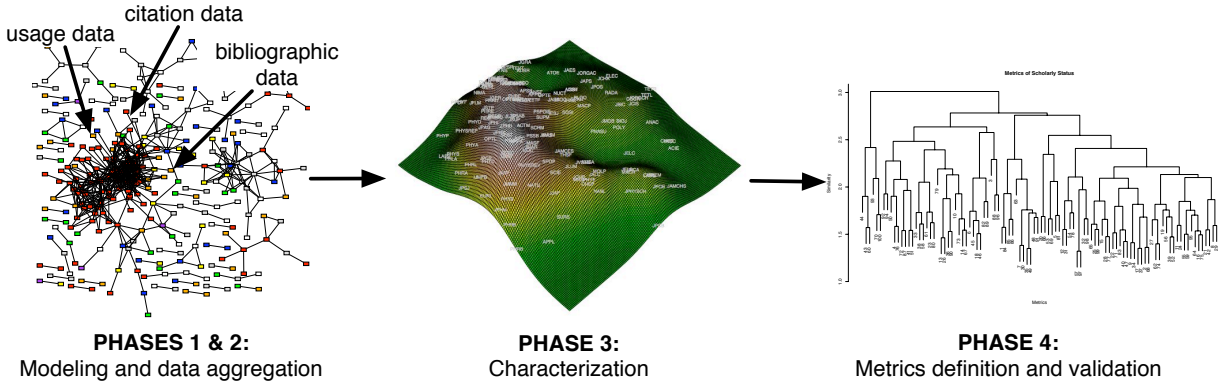

Figure 1: Overview of MESUR project phases.

It is therefore difficult to generalize any findings based on such community-specific usage data to the general scholarly community [5].

3. Cross-validation: the lack of data sets that combine usage data with other sources of information limits the possibilities for the cross-validation of usage-based metrics of scholarly impact.
4. Metrics: an enumerable number of possible indicators of scholarly impact can be defined on the basis of usage data. Which are most accurate, reliable and appropriate for a particular community and domain?

An analogy to this situation existed in the domain of Information Retrieval where a myriad retrieval algorithms could be applied to a variety of different text data sets, making it impossible to generalize the actual performance of an algorithm beyond the particular data set to which it had been applied. The lack of a standardized framework to define and validate various retrieval algorithms hampered an objective comparison of their effectiveness, and consequently their acceptance and transfer into the marketplace. This issue was addressed by the TREC [22] conferences on three levels:

1. Definition of standardized text data sets and related retrieval tasks.
2. Competition of various retrieval algorithms on the basis of (1).
3. Standardized evaluation and comparison of algorithm performance as observed in (2).

The TREC conferences have played a vital role in the development of the information retrieval domain. They have fostered innovation by offering a standardized platform for the definition and assessment of various retrieval algorithms. The TREC model has been successfully applied in other areas such as knowledge discovery and datamining [14], psychology and social science.
The MESUR[1] project, a two-year effort funded by the Andrew W. Mellon Foundation and executed by the Digital Library Research and Prototyping Team at the Los Alamos National Laboratory (LANL) Research Library (RL), seeks to perform a similar function for the development of usage-based metrics of scholarly impact, although it deviates from TREC's open competition model by implementing an *internal* program for the definition and validation of usage-based metrics[2]. The MESUR project will proceed according to the following project phases as depicted in Fig. 1:

1. **A model of the scholarly communication process:** The MESUR project will define a model of the scholarly communication process, represented as an ontology, which formally specifies the relationships between a variety of often separately represented scholarly information sources, e.g. usage, citation and bibliographic data.
2. **Creation of a reference data set:** Using the created ontology as the integrative framework, the MESUR project will aggregate and organize large-scale scholarly data into a fine-granularity semantic network that will serve as a reference data set for subsequent investigations of usage based metrics.
3. **Characterization:** Since this constitutes one of the first attempts to create a scholarly reference data set at this scale and granularity, the MESUR project will study the very structure and properties of the created semantic network to deepen our knowledge of the important demarcations and structural features of the scholarly community.
4. **Metrics definition and validation:** The MESUR project will define of a wide range of usage-based metrics, and determine their validity, reliability and scholarly correlates on the basis of the created reference data set so that guidelines can be issued with regards to their semantics and appropriate applications. The availability of a large-scale reference data set allows the similarities and divergences between various metrics to be attributed to the metrics themselves rather than variations in the underlying usage data sets. Furthermore, the structure of metric correlations can be quantitatively studied to reveal the various components of scholarly impact and how to measure them most accurately, cf. studies of human personality [16].

The following sections provide an overview of the MESUR project's workplan and architecture. First we outline the basic principles and architecture applied in our construction of

[1]Pronounced "measure", an acronym for "Metrics from Scholarly Usage of Resources".

[2]However, after that initial period 2 year period, subsequent projects may aim to achieve a platform for public competition

a model of the scholarly community process and the resulting reference data set. Second, we discuss the analysis of the properties and characteristics of this reference data set. Third, we detail the subsequent program for the definition and validation of usage-based metrics.

2. A MODEL OF THE SCHOLARLY COMMUNICATION PROCESS.

The construction of a reference data set that combines usage, citation and bibliographic data needs to be guided by an understanding of how these data sources are related in the scholarly communication process. However, to our knowledge no comprehensive ontology of the scholarly communication process exists that formally describes the relationships between usage, bibliographic and citation data, and can as such guide the creation of a large-scale reference data set from a variety of sources. Most existing scholarly ontologies separately represent such information and are focused on particular problem domains. A number of related efforts can however be identified. The ScholOnto project, [41, 40], proposes an ontology to represent scholarly documents but focuses on their argumentative, narrative structure. The ABC ontology [28], originally proposed as a conceptual model for the interoperability of metadata ontologies, represents a range of scholarly entities in an event-driven framework which has partly inspired our own efforts. Goncalves (2002)[18] proposes an XML standard for the representation of digital library usage logs, but this schema does not model the full-range of semantic relationships between usage, citation, and bibliographic data. The Web Scholars project [23] defined an ontology to represent scholarly entities [3] but does not focus on usage data. The Kendra base project, aimed at systems for the collaborative development of ontologies, contains a publication ontology[4] as well but it is largely focused on bibliographic information.

To construct the mentioned reference data set, the relationships between the various scholarly information sources need to be formally modeled in more detail. The MESUR project has therefore defined an OWL ontology [27] that can meaningfully integrate usage data with citation, bibliographic, and temporal data collected from a variety of separate sources, and does so in a form that allows it to function as the formal foundation upon on which research and development in the area of usage-based evaluation metrics can take place.

The MESUR ontology is based on two principles: representing the relationships between events and entities involved in usage, citation and bibliographic data at a fine level of granularity, but to restrict the scope and detail of such representations to data that can be realistically obtained, stored and accessed. For example, although the physical location and actual name of a user could be essential data points when one seeks to fully represent the notion of a user, such data is desirable nor practical to obtain due to privacy restrictions. The MESUR ontology therefore does not include such data entities; it is a "realistic" ontology which seeks to provide a practical, rather than an exhaustive framework for representing the scholarly communication process.

The MESUR ontology which is presently in its initial stages of development was inspired by the basic concepts underlying the OntologyX[5] framework developed by `Rights.com`. Although the latter is a proprietary, largely unpublished ontology, it is based on common principles in the domain of ontology development in particular the notion of event-based n-ary relationships called "contexts". Fig. 2 displays the basic structure of the MESUR ontology which acknowledges three abstract entities, namely:

1. **`Agent`**: authors, users, institutions, etc.
2. **`Document`**: articles, journals, conference proceedings, books, etc.
3. **`Context`**: "Uses", "Citation", "Metric", "CoAuthors", etc.

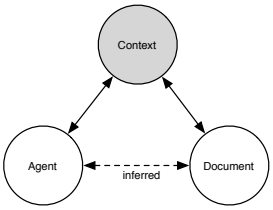

Figure 2: Notion of `Context` in MESUR ontology.

The `Context` entity is a crucial component of how the MESUR ontology represents usage, bibliographic and citation information in terms of N-ary relationships ; a `Context` ties `Agent` and `Document` entities together to express the particular nature of their relationship. Contexts are separated into two abstract types, namely `Events` and `States`. An `Event` expresses the timed action of an `Agent` upon a `Document`, whereas a `State` expresses a persistent, static state of a `Document` or `Agent`.

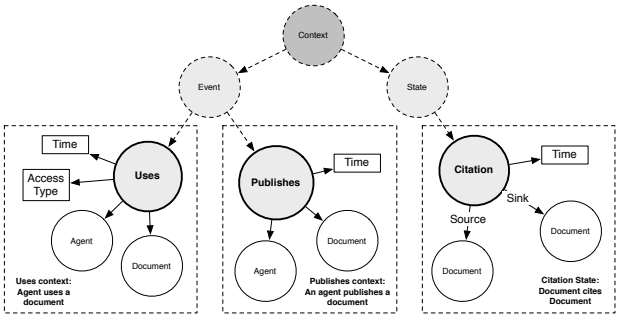

Figure 3: Use of Context class in MESUR ontology to express usage, citation and publishing relationships.

The following relationships (and many others) can be represented using this framework:

- An **author** (`Agent`) **publishes** (`Context:Event`) an **article** (`Document`)
- A **user** (`Agent`) **uses** (`Context:Event`) a **journal** (`Document`)

[3] `http://www.kampa.org/phd/ontology/s_community.html`
[4] `http://base2.kendra.org.uk/main_page`
[5] `http://www.ontologyx.com/`

- An **article** (`Document`) **cites** (`Context:State`) another **article** (`Document`
- An **author** (`Agent`) is a **Coauthor** (`Context:State`) with another **author** (`Agent`)
- A **journal** (`Document`) has an Impact Factor value (`Context:State`) of 0.75 (`literal`).

Fig. 3 provides a visual example of how different contexts can express different document-agent relationships, but omits the required OWL terminology and specifications.
Bibliographic and demographic information are stored as the properties of instances of the `Document` and `Agent` classes respectively. The focus in the latter cases again is on the data that is realistically obtainable and required to achieve MESUR project objectives, rather than a theoretically exhaustive description of either `Document` or `Agent` entities. The present version of the MESUR ontology can be accessed at `http://www.mesur.org/schemas/2007-01/mesur/` and is continuously under development as it is being used to aggregate and normalize data obtained from a variety of resources.

3. DATA AGGREGATION

After establishing a model of the scholarly communication process actual data needs to be obtained and stored to create the mentioned reference data set that will serve as the basis of a program for the definition and validation of usage-based metrics. This reference data set needs to be of sufficient scale to be representative of the scholarly communication process. In this section we discuss our efforts to achieve and store such a reference data set.

3.1 Sampling and scaling

The realization that studies of usage-based scholarly evaluation metrics require data sets of sufficient scale and granularity has prompted a number of efforts to aggregate and combine usage data with other scholarly information. Bollen et al (2005)[3] propose an architecture for the large-scale aggregation of usage data recorded by linking servers, encoded as OpenURL ContextObjects and aggregated by means of the OAI-PMH standard. The COUNTER project[6] [39] has established a standard framework for expressing large-scale usage statistics of scholarly journals as recorded by on-line publisher services. The SUSHI project[7] aims at making COUNTER data harvestable. The Interoperable Repository Statistics[8] project looks into the definition of interoperable usage statistics for OAI-PMH-compliant repositories (institutional and discipline-based). The CiteBase project[9] extracts citations from on-line, freely available publications and records its local web site hits to offer services that contrast the resulting usage data to its collections citation rates. However, none of the mentioned projects have implemented an integration of usage, citation and bibliographic data across a manifold of institutions and at the scale and granularity that characterizes the MESUR project. For example, the LANL RL has aggregated one year of linking server data collected from 23 California State University (CSU) campuses. This data concerns a total of 167,204 users and 2,133,556 documents, and can be considered one of the more diverse and detailed usage data sets that have been obtained for research purposes. However, the CSU data usage data is not formally integrated with any citation or bibliographic data, and does not represent a significant sample of the scholarly community. In comparison with this initiative, the MESUR project seeks to aggregate usage, citation and bibliographic data for:

1. At least 50 million documents, including all associated metadata and references.
2. Approximately 70 million users and authors combined.
3. Approximately 1 billion usage events and 500 million citations.
4. A significant sample of the world's major publishers and scholarly institutions including the largest library consortia.

According to the above described ontology, this number of scholarly entities will result in about 3 to 5 billion semantic relationship statements, i.e. (Subject, Predicate, Object) triples, pertaining to a variety of combinations of usage, citation or publication events. Given the large scale of MESUR's intended reference data set, two equally important issues need to be addressed. First, agreements need to be achieved with usage data providers to furnish usage data in sufficient quantities to achieve a significant sample of the scholarly community. Second, an architectural framework needs to be created to aggregate, normalize and store scholarly data at such scale.
These two issues will be discussed in the following sections beginning with an overview of MESUR efforts to obtain usage from a variety of providers and detailing the project's particular data acquisition, parsing and storage approaches.

3.2 Project Collaborators and partners.

The MESUR project is critically dependent on usage data that is provided by an as wide an array of providers and institutions as possible, to achieve a reference data set that is maximally representative of the scholarly community. The MESUR project therefore contacted a wide range of potential providers of usage data and provided them with a checklist of requested and desired data items as well as a statement with regards to MESUR's privacy guidelines. Table 1 thematically summarizes the requested data items and privacy guidelines, the latter of which apply to all stakeholders involved in the recording of usage data.

MESUR has at this point achieved collaborative agreements with a large number of parties, including a set of university library consortia, national library service providers and international publishers. Furthermore, the COUNTER Executive Committee has expressed its willingness to provide community support and collaborate with MESUR on issues relevant to the assessment of scholarly impact such as extensions to the COUNTER Code of Practice and examining the possibility of a Usage Impact Factor.

3.3 Aggregation, filtering, parsing, and deduplication

Processing raw usage data from a variety of sources entails four distinct issues which the MESUR project acknowledges in its infrastructure:

[6] `http://www.projectcounter.org/`
[7] `http://www.niso.org/committees/SUSHI/SUSHI_comm.html`
[8] `http://irs.eprints.org/`
[9] `http://www.citebase.org/`

Data requirements.

Data	Requirement
Granularity	Ability to separate individual events and types of events pertaining to specific documents.
Agent or user identifier	Identifier of agent, user or session, e.g. session identifier, user login, IP address or other.
Document identifiers	Identifier of object which is subject of event, e.g. DOI, PMID, URI, local ID, or other.
Event time	Ability to extract date and time for individual events at second, minute or hour granularity.

Stakeholder privacy.

Stakeholder	MESUR privacy guideline
Individual user	User identity will never be revealed. Usage data can be anonymized by provider or MESUR.
Institutional Privacy	Institutional identity will only be revealed at consent of institution.
Usage data provider	Provider identify will only be revealed at consent of provider.

Table 1: Summarized data requirements checklist and privacy guidelines for potential usage data providers

1. Aggregation and large-scale data management.
2. Filtering, i.e. the removal of automatically generated usage events, e.g. those caused by crawlers, spiders and bots.
3. De-duplication, i.e. the identification of events that pertain to identical documents.
4. Uncertainty quantification, i.e. determining the validity and noise-levels of obtained usage data.

The MESUR project acknowledges that there exists a multitude of different methods to record usage and that potential partners of the project will each have adopted a particular configuration of methods, systems and formats. The aggregation of usage data from a variety of sources, e.g. web servers, link resolvers, and specialized databases applications, represents a considerable challenge. A large proportion of the MESUR project's resources is therefore dedicated to this task in order to reduce the effort required of individual usage data providers and thus reduce the threshold for their participation.

As shown in Fig. 4, the project architecture is geared towards accepting raw usage data from a variety of sources such as individual libraries (or consortia) and publishers. After a screening process to determine whether the particular usage data contains the MESUR-required data items (see Table 1), the usage data sets are transfered from the usage data providers by any provider-preferred means, e.g. FTP, secure transfer, or physical media such as DVD. Since user and institutional privacy are important issues, usage data is always anonymized, either in advance by the provider or by the MESUR project before processing commences. Specific privacy guidelines apply to all levels of this process (see Table 1).

In addition to anonymization, the MESUR project also records data provenance information at this processing stage. The logistics involved with the management of large-scale usage data sets from a variety of providers represents a major challenge. The origin, manipulation and quality of each usage data set needs to be carefully documented to ensure future verification of MESUR results. The MESUR project has therefore developed an XML Schema to represent origin, lineage, and data quality information, so that an XML formatted data file can accompany all provided usage data sets to record all project relevant details with regards to its provenance. An overview of this schema is shown in Fig. 5. The schema, which is still under development, is published on the MESUR web site:

`http://www.mesur.org/schemas/2007-01/provenance.xsd`.

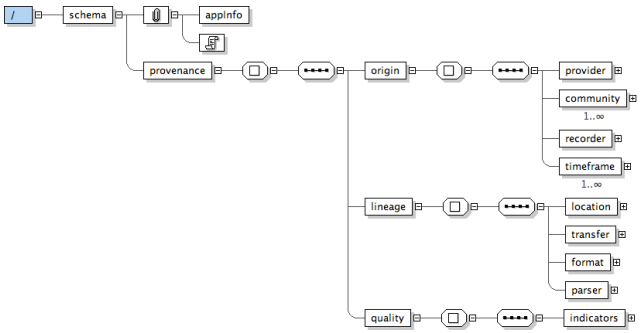

Figure 5: Summarized XML schema to represent usage data provenance, lineage and data quality information.

After the recording of provenance data, the raw usage data is filtered and parsed by a set of provider-specific parsers. The filtering process entails the detection and exclusion of usage events which have been caused by robots, crawlers, spiders and other non-human agents. Detection of such spurious events is conducted by taking into account a variety of statistical parameters associated with the online behaviors that characterizes non-human agents. Event exclusion rates in the filtering process in addition provide an indication of the reliability of the particular data set; the number of non-human agents increases the uncertainty by which the log data expresses valid usage. For that reason, data quality metrics such as event exclusion rates can be added to the usage data's provenance information at this phase.

In addition to being parsed and filtered, the provided usage data must also be de-duplicated, i.e. usage events pertaining to the same document must be identified to avoid duplicate entries. Clearly this is more an issue for article-level usage data than it is for journal-level usage data. The MESUR project leverages a de-duplication system that was deployed in earlier projects. It relies on a bibliographic masterlist of more than 50 million documents on the basis of which a set of Bayesian algorithms attempts to identify an event's document on the basis of its identifiers and metadata. If an event's document matches an existing item in the masterlist it is assigned a pre-existing internal identifier. If not,

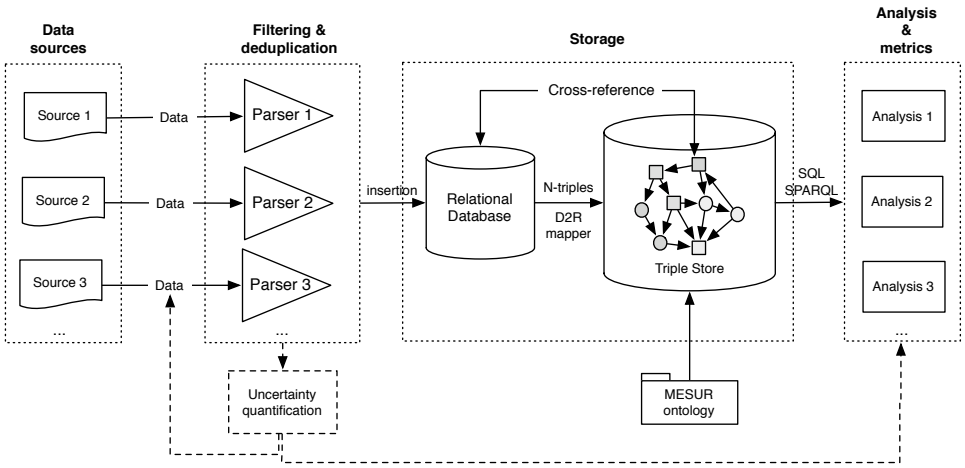

Figure 4: General data flow in MESUR project.

it is assigned a new identifier and added to the masterlist itself. As such, the document of any subsequent event will be matched to a continuously expanding bibliographic masterlist.

A consequence of this approach is that usage data will be most successfully de-duplicated for those documents congruent to the pre-established masterlist. This situation is represented in Fig. 6; many documents for which usage data was recorded will either not be included in the masterlist or lack sufficient metadata and identifiers to perform a sufficiently accurate match; they will thus be de-duplicated less reliably. In spite of the lower reliability of these con-congruent documents, they may still play an important role in subsequent investigations.

The MESUR project acknowledges that filtering and de-duplicating raw usage data are inherently statistical procedures. De-duplication and filtering success rates will therefore need to be included in assessments of data quality; these will in their turn inform the reliability attributed to the outcomes of any subsequent data analysis. The latter is a component of the project's work plan referred to as "uncertainty quantification". During the processing of MESUR's data, the project will record a variety of indicators that express the degree to which each successive stage of data processing is subject to statistical errors, i.e. uncertainty. This information will be recorded according to the mentioned provenance schema (see Fig. 5). When finally usage-based metrics are defined and validated in the later stages of the project, these uncertainty indicators will be compounded into overall indicators that express the reliability that can be attributed to any of the project's conclusions or guidelines.

3.4 Storage architecture

After aggregation, filtering, parsing, and de-duplication, the resulting usage data needs to be stored for future access. Two competing paradigms can be considered for the storage of semantic information consisting of a mix of citation, bibliographic and usage data, namely the use of relational databases and semantic triple stores. The former are a well-established paradigm. However, a number of standards and tools have emerged in recent years to encode ontologies and manage conforming large-scale data sets consisting of

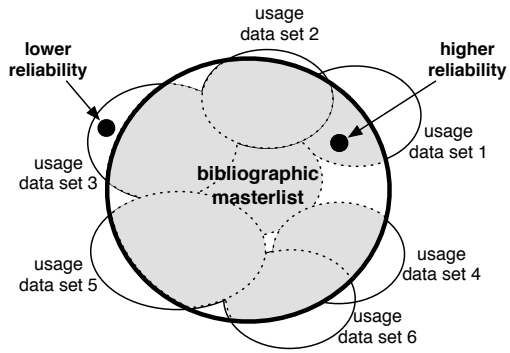

Figure 6: Using a bibliographic reference data set to normalize and de-duplicate usage data.

`(object, predicate, subject)` triples much like relational databases presently store tabulated data:

1. The W3C Resource Description Framework (RDF[10]) and Web Ontology Language (OWL[11]) provide XML standards for the representation of ontologies and semantic networks.
2. Triple store databases: database systems that efficiently load, manipulate, and query large-scale RDF triple data sets , e.g. Franz's AllegroGraph[12], Kowari[13] and Oracle 10g + Spatial[14].
3. Access interfaces: Triple store query languages such as RDQL (Jenna), iTQL (Kowari), Redland (3store), SeRQL (Sesame), SPARQL and customized front-ends based on Prolog and Lisp which enable convenient access to the stored sets of triples.

Triple store databases have the advantage that they are geared towards storing RDF triples [31] and allow semantic

[10] `http://www.w3.org/TR/rdf-concepts/`
[11] `http://www.w3.org/TR/owl-features/`
[12] `http://www.franz.com/products/allegrograph/`
[13] `http://www.kowari.org/`
[14] `http://www.oracle.com/technology/tech/semantic_technologies/index.html`

inferencing on their content. Furthermore, they offer much greater flexibility in terms of data modeling than relational database do. On the other hand, their scalability and retrieval efficiency are generally not on a par with the most competitive relational database products.

The MESUR project has decided to leverage a combination of relational database and triple store approaches to optimize scalability and inference speed. The MESUR relational database is used to store all usage information for parsing and de-duplication, in conjunction with all document and agent metadata. Using the D2R Mapper [1], a set of N-triples is generated from the relational database according to MESUR's ontology. These N-triples pertain only to semantic information in which particular ontological entities are identified by means of unique URI's. No actual agent, document or other metadata is included except where they are expected to be an essential part of future metrics investigations, e.g. agent affiliation data and metric values. The resulting N-triples are then loaded into the triple store. For example, the triple store may contain a triple stating that the `Agent` identified by the URI `http://www.mesur.org/#1` "Uses" the `Document` identified by the URI `http://www.mesur.org/#2`, i.e. a total of 3 URIs (including the "Uses" context URI), but the actual document identifier, document title, document author, user affiliation, etc. would only be stored in the project's relational database. Any URI will thus need to be dereferenced from the relational database when necessary, as shown in Fig. 4. This approach will increase the overall performance and scalability of the triple store by not burdening it with unnecessary metadata. In addition, most inference algorithms are not expected to require full item metadata and entity URI's can be dereferenced from the relational database after inferencing has been completed.

4. CHARACTERIZATION

After the data aggregation and normalization phase, the MESUR project will be equipped with a large-scale reference data set that combines a multitude of primary and derived bibliographic, citation, usage and temporal relationships in a triple store database consisting of 3 to 5 billion triples. In this phase of the MESUR project, a lay of the land will be investigated, i.e. the general properties of the scholarly communication process, so that it can guide the following project phase that is focused on the definition and validation of usage-based scholarly evaluation metrics.

There exists an impressive body of literature on the subject of characterizing the scholarly communication process, but most of these studies remain in the area of theoretical scholarly research and are mainly based on citation or bibliographic data. We can distinguish two types of approaches in this domain:

1. Quantitative network metrics: pertaining to the general graph-theoretical properties of the relationships in the scholarly domain.
2. Qualitative descriptions: explorative clustering and spatial maps intended to elucidate particular structural features.

The first approach is strongly grounded in social network analysis and has proven its value in a range of domains such as studies of the World Wide Webs structure. Its objective is to describe the general structural properties of networks in terms of their topological features. For example, Newman (2001) [34, 35] examines the degree of clustering in co-authorship and citation networks, and the presence of small-world features in co-authorship networks. Jeong (2003)[20] demonstrates the presence of preferential attachment in citation networks. Kretschmer et al. (2003) [24] determine the cohesion of citation networks.

The second approach is largely geared towards producing human-readable, explorative representations of the structure of citation data. For example, Boyak (2002,2004,2005) [10, 8, 9] produces three-dimensional maps of the scholarly community on the basis of citation data using the VxInsight tool developed at Sandia Laboratories. The mappings are generated for different time periods revealing scholarly trends and emerging research fronts. Leyesdorff (2004,2006) [29, 30] uses a principal component analysis to cluster and organize journals on the basis of the Journal Citation Records. Chen (2001) maps the intellectual structure of knowledge. Nagpaul (2002)[32] visualizes the structure of cooperation networks among Indian institutions.

However, few attempts have been made to apply the mentioned analysis to usage data, or to compare the properties of the scholarly communication process across the different layers of bibliographic, citation and usage data. In one of the few publications in the latter area, Bollen (2006)[4] mapped the structure of the LANL user community by applying a principal component analysis to journal relationship matrices derived from LANL usage data. The resulting maps were compared to those generated on the basis of citation data, thereby revealing research communities at LANL whose characteristics differed from those of the general scholarly community.

The MESUR project's characterization phase will thus take place along three lines of explorative research.

First, data will be correlated across a range of data types in the created semantic network to validate its content. Usage rates can for example be correlated to citation rates on the level of authors, articles, journals and particular affiliations. Discrepancies may point to data errors or genuine phenomena in the scholarly community. A regression analysis can reveal particular correlations between various primary and derived relationships, such as how particular scholarly domains, potentially derived from bibliographic data, influence usage and citation rates. The nature of the generated semantic network will allow us to separate particular subsets and investigate their relationships to other subsets.

Second, the project will aim to determine the topological properties of the generated semantic network across all layers of bibliographic, citation and usage data. This second line of research seeks to quantify the overall properties of the generated semantic network using a set of social network metrics, in particular focusing on the following indicators:

1. Topological parameters: Cohesiveness, density, clustering degree, cliquishness, etc.
2. Small world features: Degree distributions, network diameter, connected sub-components, etc.

The results of this investigation will serve to further validate and characterize the generated semantic network by comparing its features to those of other data sets such as e.g. the WWW.

The third line of research seeks to explore the rich set of relationships embedded in the generated model of the scholarly communication process by mapping and visualizing its structure including the longitudinal aspects of scholarly trends. The following methods can be pursued:

1. Cluster analysis: Hierarchical and k-means on the basis of author, article, journal and affiliation data.
2. Dimensionality reduction techniques: Principal component analysis [21] or multi-dimensional scaling of author, article, journal and affiliation relationships.
3. Longitudinal analysis: Visual mapping of domain changes over time.

Each proposed line of research in this phase is intended as an explorative effort aimed at determining the overall properties of the generated semantic network and the degree to which the results of such an investigation can inform the subsequent definition of usage-based scholarly evaluation metrics. This project phase is expected to encounter severe scalability issues. Few statistical analysis packages have been designed to analyze data sets of the size expected of the generated semantic network. Existing collaborations with the Theoretical Physics and Computing divisions at the Los Alamos National Laboratory will be crucial to successfully complete this phase of the project.

5. METRICS DEFINITION AND VALIDATION

The results of the projects previous phase will provide a first survey of the properties of the scholarly communication process that will in turn provide the essential groundwork for the investigation of viable usage-based metrics.
A myriad of different usage-based scholarly evaluation metrics can in principle be proposed, but the issues at hand are first *whether* and second *how* any metric actually reflects scholarly impact. The latter two issues can be framed in terms of metric validity and semantics:

Metric validity Does the metric actually indicate what it is intended to indicate, namely scholarly impact?

Metric semantics If so, what aspect or facet of impact does the metric indicate, e.g. prestige, popularity, notoriety, etc?

The MESUR project aims to address these issues as depicted in Fig. 7. First, a wide range of plausible metrics of scholarly impact will be defined which will both adopt and extend the set of metrics presently proposed in the literature, e.g. the Reading Factor [15], Usage Impact Factor [5] and the citation-derived ISI journal Impact Factor [17], as well as more structural indicators of impact, such as the most common social network centrality metrics [6, 2] applied to a combination of usage, bibliographic, and citation data. Hybrid metrics of impact that rely on multiple scholarly information sources have found little attention in the literature; the MESUR project will therefore be particularly active in this area. For example, hybrid metrics of impact could be defined that take into account a combination of usage and citation rates in addition to inferences based on co-authorship data. The created reference data set will be instrumental in the definition of such inference-based metrics. Indicators of scholarly impact will be applied to a range of scholarly entities such as documents, journals, authors as well as institutions.
A short-list of the classes of usage-based metrics of scholarly impact that will be considered in the MESUR project is listed below.

1. Raw and normalized usage frequencies.
2. PageRank [37, 12] applied to usage networks.
3. H-index [19, 11] calculated for journals and authors.
4. Social network centrality metrics calculated over a range of bibliographic, citation and usage-derived relationships [42, 6].
5. Hybrid metrics such as the Y-factor [2].
6. Inference-based metrics of scholarly impact relying on rule-based definitions of impact [38].

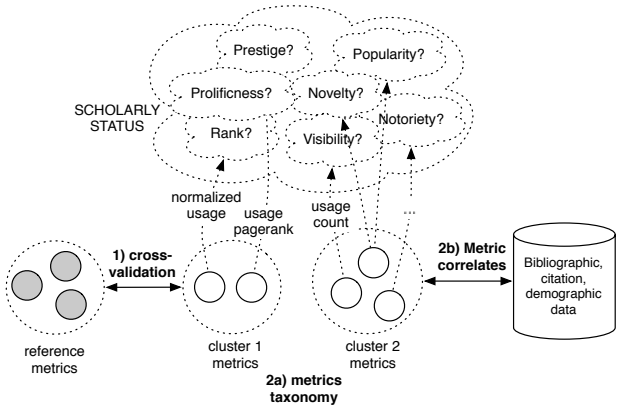

Figure 7: Cross-validating usage-based metrics.

The second stage of this process entails the issue of metric validity and is addressed by a program of metric cross-validation. This requires the use of reference metrics which have in the past proven their ability to indicate particular aspects of scholarly impact. MESUR is considering using raw article citation counts, as well as the ISI journal Impact Factors and COUNTER statistics for this purpose. The latter in particular will play an important role in the validation of usage-based metrics since it is based on publisher-generated usage statistics. The ability of any particular metric to indicate scholarly impact, can be then assessed by correlating its outcomes to those of the established reference metrics.
Third, it is understood that each validated metric will indicate a different, but related aspect of scholarly impact. The MESUR project will therefore seek to determine the semantics of the defined and validated metrics. This problem will be tackled by a variety of methods. Correlating the outcomes of all validated metrics will reveal clusters of metrics which indicate a similar facet of scholarly impact. In fact, a taxonomy of usage-based metrics can be derived from performing a factor- and cluster-analysis on the correlation structure of the different metrics, similar to the research that led to the formulation of the Big Five model of human personality [16]. Subsequently, the outcomes of groups of metrics can be correlated to citation and bibliographic data to interpret their semantics. For example, a usage-based metric whose results strongly correlate with the ISI journal Impact Factor and the status of authors' alma maters could

be interpreted to indicate scholarly impact.
On the basis of this metric cross-validation and interpretation program, the MESUR project will issue guidelines with regards to the validity and semantics of a large set of usage-based metrics, their appropriate application and the degree to which they approach aspects of existing indicators such as the ISI journal Impact Factor.

6. TIMELINE AND DELIVERABLES

As identified in the introduction the two-year MESUR project will proceed according to four distinct phases: definition of a model of the scholarly communication process (phase 1), aggregation of large-scale usage, citation and bibliographic data into a reference data set according to this model (phase 2), the characterization of the scholarly community on the basis of the created reference data set (phase 3) and finally a program for the definition and validation of usage-based metrics of scholarly impact (phase 4), finalized by the issuing of a set of guidelines with regards to applications of usage-based metrics of scholarly impact. The project's timeline and deliverables are summarized in Fig. 8.

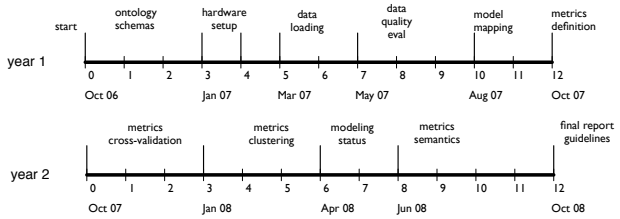

Figure 8: Summarized two-year timeline for MESUR project.

7. CONCLUSION

Usage data is expected to play an increasingly important role in the assessment of scholarly impact. However, the acceptance of usage-based metrics for scholarly evaluation is lagging in spite of their considerable potential. This can be attributed to a number of factors identified in the introduction of this paper, but most importantly the lack of a standardized platform upon which a program for the identification and validation of a range of usage-based metrics can be founded. Indeed, most investigations of usage-based metrics of scholarly impact focus on single metrics whose characteristics are explored on the basis of usage data sets relevant only to particular scholarly communities.
The MESUR project addresses this issue on multiple levels. First, the establishment of a model of the scholarly communication process, i.e. an ontology relating citation, bibliographic and usage data, serves as the formal groundwork to organize the multiple relationships between the products of the scholarly community. Second, on the basis of the defined ontology a large-scale reference data set is created upon which investigations of usage-based metrics can be conducted. This reference data set combines usage data from a wide range of institutions to provide a representative sample of the scholarly community. By its relations to citation and bibliographic data, it can be validated and enriched. Third, the MESUR project will study the properties of this reference data set to determine the major demarcations and clusters of practice in the scholarly community, a factor which is known to greatly affect assessments of impact. Fourth, a multitude of usage-based metrics of scholarly impact will be defined and validated against the established reference data set. By doing so on a common platform, it is possible to quantitatively evaluate the semantic and relationships of various metrics, thereby establishing an understanding of which aspects of impact they represent and where they can or should be applied. We believe the MESUR project's role will not be limited to its final objective, namely issuing a set of guidelines with regards to usage-based metrics; by establishing a large-scale, semantic reference data set it provides the foundation for a continued development of this domain, and the possibility of a more accurate and complete assessment of scholarly impact in the future.

8. ACKNOWLEDGMENTS

This research is supported by a grant from the Andrew W. Mellon Foundation.

9. REFERENCES

[1] C. Bizer. D2R - a database to RDF mapping language. In *The Twelth International World Wide Web Conference (WWW03)*, Budapest, Hungary, May 2003.

[2] J. Bollen, M. A. Rodriguez, and H. Van de Sompel. Journal status. *Scientometrics*, 69, 2006.

[3] J. Bollen and H. Van de Sompel. An architecture for the aggregation and analysis of scholarly usage data. In *Joint Conference on Digital Libraries (JCDL2006)*, pages 298–307, Chapel Hill, NC, June 2006.

[4] J. Bollen and H. Van de Sompel. Mapping the structure of science through usage. *Scientometrics*, 69(2), 2006.

[5] J. Bollen and H. Van de Sompel. Usage impact factor: the effects of sample characteristics on usage-based impact metrics. *arxiv.org*, cs.DL/0610154, 2007.

[6] J. Bollen, H. Van de Sompel, J. Smith, and R. Luce. Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management*, 41(6):1419–1440, 2005.

[7] M. Bordons, M. T. Fernandez, and I. Gomez. Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics*, 53(2):195–206, 2002.

[8] K. W. Boyack. Mapping knowledge domains: Characterizing PNAS. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1), April 6, 2004.

[9] K. W. Boyack, R. Klavans, and K. Boerner. Mapping the backbone of science. *Scientometrics*, In press, 2005.

[10] K. W. Boyack, B. N. Wylie, and G. S. Davidson. Domain visualization using vxinsight(r) for science and technology management. *Journal of the American Society for Information Science and Technology*, 53(9):764–774, 2002.

[11] T. Braun, W. Glaenzel, and A. Schubert. A hirsch-type index for journals. *The Scientist*, 19(22):8, November 2005.

[12] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[13] T. Brody and S. Harnad. Earlier web usage statistics as predictors of later citation impact. Eprint 10647, University of Southampton, ECS, Intelligence, Agents, and Multimedia Group, 2005.

[14] R. Caruana, T. Joachims, and L. Backstrom. KDD-Cup 2004: Results and analysis. In *Proceedings of the Tenth Annual ACM Knowledge Discovery and Datamining Conference (KDD2004)*. ACM Press, August 2004.

[15] S. J. Darmoni, F. Roussel, J. Benichou, B. Thirion, and N. Pinhas. Reading factor: a new bibliometric criterion for managing digital libraries. *Journal of the Medical Library Association*, 90(3):323–327, 2002.

[16] J. M. Digman. Higher-order factors of the Big Five. *Journal of personality and social psychology*, 73(6):1246 – 1256, 1997.

[17] E. Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. John Wiley and Sons, New York, 1979.

[18] M. A. Goncalves, M. Luo, R. Shen, M. F. Ali, and E. A. Fox. An XML log standard and tool for digital library logging analysis. In M. Agosti and C. Thanos, editors, *ECDL 2002: LNCS 2458*, pages 129–143, Berlin, September 2002. Springer-Verlag.

[19] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

[20] H. Jeong, Z. Neda, and A. L. Barabasi. Measuring preferential attachment in evolving networks. *Europhys. Lett.*, 61(4):567–572, 2003.

[21] I. Jolliffe. *Principal Component Analysis*. Springer, New York, 2002.

[22] K. S. Jones. Further reflections on trec. *Information processing and management*, 36(1):37 – 85, 2000.

[23] S. Kampa and L. Carr. Web scholars (poster). In *Nineth International WWW Conference*, pages 44–45, Amsterdam, NL, May 2000.

[24] H. Kretschmer and R. Rousseau. A measure for the cohesion of weighted networks. *Journal of the American Society for information science*, 54(3):193–203, 2003.

[25] M. J. Kurtz, G. Eichhorn, A. Accomazzi, C. S. Grant, M. Demleitner, and S. S. Murray. The bibliometric properties of article readership information. *JASIST*, 56(2):111–128, 2004.

[26] M. J. Kurtz, G. Eichhorn, A. Accomazzi, C. S. Grant, M. Demleitner, and S. S. Murray. Worldwide use and impact of the NASA Astrophysics Data System digital library. *JASIST*, 56(1):36–45, 2004.

[27] L. W. Lacy. *OWL: Representing Information Using the Web Ontology Language*. Trafford Publishing, 2005.

[28] C. Lagoze and J. Hunter. The ABC Ontology and Model. *Journal of Digital Information*, 2, 2001.

[29] L. Leydesdorff. Clusters and maps of science journals based on bi-connected graphs in journal citation reports. *The journal of documentation*, 60(4):371–427, 2004.

[30] T. Leydesdorff. Can scientific journals be classified in terms of aggregated journal-journal citation relations using the journal citation reports? *Journal of the American Society for Information Science and Technology*, 57(5):601 – 13, 2006.

[31] E. Miller. An introduction to the resource description framework. *D-Lib Magazine*, May 1998.

[32] P. S. Nagpaul. Visualizing cooperation networks of elite institutions in india. *Scientometrics*, 54(2):213–228, 2002.

[33] Nature. Editorial: Not-so-deep impact. *Nature*, 435:1003–1004, June 2005.

[34] M. E. J. Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):025102+, 2001.

[35] M. E. J. Newman. Scientific collaboration networks. II. clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102+, 2001.

[36] T. Opthof. Sense and nonsense about the impact factor. *Cardiovascular Research*, 33:1–7, 1997.

[37] G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications: theory, with application to the literature of physics. *Information processing and management*, 12(5):297 – 312, 1976.

[38] M. A. Rodriguez. Grammar-based random walkers in semantic networks. `http://tinyurl.com/278jtf`, 2007.

[39] P. T. Shepherd. Project COUNTER - Setting international standards for online usage statistics. *Journal of Information Processing and Management*, 47(4):245 – 257, 2004.

[40] S. B. Shum, J. Domingue, and E. Motta. Scholarly discourse as computable structure. In *Second International Workshop on Structural Computing - ACM HT2000*, pages 120–128, San Antonio, TX, June 2000.

[41] S. B. Shum, E. Motta, and J. Domingue. ScholOnto: an ontology-based digital library server for research documents and discourse. *International Journal on Digital Libraries*, 3(3):237–248, 2000.

[42] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK, 1994.