

# Enriching Knowledge Domain Visualizations: Analysis of a Record Linkage and Information Fusion Approach to Citation Data

Marie B. Synnestvedt, MEd<sup>1,2</sup>

<sup>1</sup>Drexel University College of Information Science and Technology, Philadelphia PA

<sup>2</sup>University of Pennsylvania School of Medicine, Philadelphia PA

## Abstract

*This article presents a study of the use of data preparation for data mining methodology to prepare biomedical citation data for visualization. Deterministic record linkage models were compared with probabilistic record linkage in a situation for which the truth is known through the use of gold standard or truth datasets. The linkages are evaluated on data from the Web of Science (WOS) and Medline citation databases. Sensitivity, specificity, and overall performance of record linkage models were empirically compared with ROC analysis. Data quality and visualization metrics are presented for datasets prepared with and without probabilistic record linkage and information fusion of Medline abstracts and MESH terms into WOS citation records. The major contributions of this work are to specifically develop a novel model of record linkage for biomedical citation databases, with the objective of improving and enriching biomedical knowledge domain visualizations.*

## Introduction

There are strong parallels between citation databases and data warehouses, and a data preparation for data mining approach is applicable to the data extracted from citation databases. One of the most important, time consuming, and difficult steps in the Knowledge Discovery in Databases (KDD) process is data preparation. The advantage of data preparation is that it can substantially improve the overall quality of patterns mined<sup>1</sup>. Record linkage is a data preparation method of identifying database records that are syntactically different but refer to the same entity and lack a common unique identifier. Probabilistic record linkage methods are based on statistical theory and artificial intelligence techniques, and used to determine the probabilistic matching between records and for extracting a unique identifier or a set of variables acting as an identifier<sup>2, 3, 4, 5, 6</sup>. While many probabilistic record linkage studies can be found in the medical research literature, no work appears to exist in the library science literature on application of

this theory and methodology to preparation of citation data for subsequent analysis.

We have previously reported studies of visualization of the domain of medical informatics using CiteSpace II, a Java application which combines information visualization methods, bibliometrics, and data mining algorithms in an interactive visualization tool for extraction of patterns in citation data<sup>7, 8</sup>. Highly cited and pivotal documents, areas of specialization within a knowledge domain, and emergence of research topics are mapped for discovery through visual pattern recognition. The primary sources of data for CiteSpace analyses are the ISI Web of Science bibliographic databases, and a secondary source is the National Library of Medicine's Medline bibliographic database via the PubMed system. The two data sources must be analyzed separately. The major distinction between the two sources of data from an analytic perspective is the availability of citation rate and cited reference data from WOS, and the availability of medical subject headings (MeSH) from Medline. Citation rates and cited references are the key to identifying pivotal documents and trends, and MeSH terms are useful for organizing documents by subject content according to a controlled vocabulary that is familiar and relevant to the medical community. Detailed reports of the theoretical and methodological basis on which CiteSpace II was developed can be found in Chen, 2004<sup>9</sup> and Chen, 2005<sup>10</sup>.

One of the challenges of working with the data repositories typically used in data mining is that real world data tend to be incomplete, noisy, and inconsistent. Citation data has characteristics that fit with this description of real world data. And if there are systematic patterns of missing citation data by publication date or specialty within a knowledge domain, this could lead to a biased analysis and result in visualizations that contain misinformation. A post-study analysis of the medical informatics datasets used in the previous visualization studies revealed anomalous patterns of missing keywords and abstracts in citation data obtained from Web of

Science relative to citation data obtained from Medline (Figures 1 & 2).

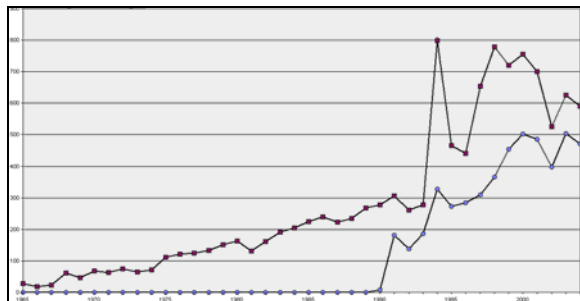


Figure 1. Number of citation records with keywords by year of publication, Medline (top line) vs. WOS.

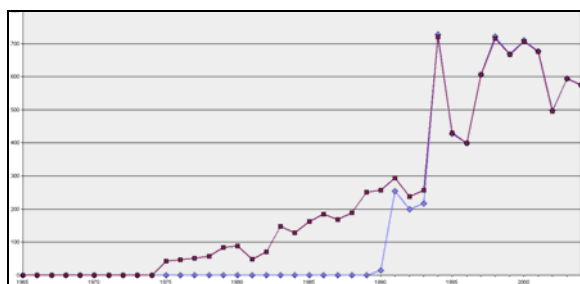


Figure 2. Number of citation records with abstracts by year of publication, Medline (top line) vs. WOS.

Additional characteristics of citation data that necessitate a record linkage approach are the absence of a unique identifier to join records and the inconsistency between databases in the recording of elements of a citation. Two recent standards for unique document identifiers that theoretically could be used to link citation records are the publisher item identifier (PII) and digital object identifier (DOI), but the adoption and availability of these identifiers is limited and varies by journal publisher and database. In a sample of 18,197 records collected from the Medline database, 27% included a PII and 9% included a DOI. Neither identifier was available in an equivalent sample collected by direct export from the WOS database. A search of the online Bluesheets documentation for the DIALOG system indicated DOI availability only in non-medical databases (primarily engineering fields), and PII availability in the SCISEARCH and SOCIAL SCISEARCH databases from June 2003 forward.

## Methods

**Data:** Datasets previously obtained from Medline and WOS and developed as a test bed for visualization of Medical Informatics are detailed in Synnestvedt et. al., 2005<sup>7</sup>. The WOS dataset of 11,752 citation records from twelve journals covered

forty years from 1964-2004. For purposes of record linkage testing, the Medline dataset was expanded for this study with data from four additional journals.

Journal Title	WOS	Medline
Artificial Intelligence In Medicine	449	485
Cin-Computers Informatics Nursing	121	101
Computer Methods And Programs In Biomedicine	1609	1584
Computer Programs In Biomedicine	437	512
Computers And Biomedical Research	1403	1418
Computers In Nursing	119	650
Ieee Transactions On Information Technology In Biomedicine	210	304
International Journal Of Bio-Medical Computing	1021	1198
International Journal Of Medical Informatics	736	718
International Journal Of Technology Assessment In Health Care	742	1351
Journal Of Biomedical Informatics	152	157
Journal Of The American Medical Informatics Association	1674*	689
Proceedings / The Annual Symposium On Computer Application [Sic] In		1009
Proceedings : A Conference Of The American Medical Informatics		329
Proceedings / Amia Annual Symposium Amia Symposium		946
Amia ... Annual Symposium Proceedings...		458
M D Computing	500*	836
Medical Decision Making	871*	1145
Medical Informatics And The Internet In Medicine	136	134
Methods Of Information In Medicine	1572*	1895
Bioinformatics (Oxford, England)		2198
Computers In Biology And Medicine		1219
Journal Of Telemedicine And Telecare		1103
Medinfo		1332
Total	11,752	21,771

Table 1. Distribution of records by Dataset and Journal. \*: Meeting abstracts excluded.

**Standardization Procedures:** Standardization of variables is necessary to increase performance of the record linkage<sup>6</sup>. All standardization routines were developed in Microsoft Access. The raw export files from the citation databases were processed to a normalized record structure, and variables parsed and standardized to the “least common denominator” format of WOS and Medline.

**Sample Selection for truth datasets and modeling:** The first step in developing the truth datasets was to randomly split the WOS dataset into equal pools of potential case and control citations from which three ten-percent samples without replacement are drawn. Truth datasets for each of these samples were validated using a combination of relational database queries and manual review of source documents.

**Record Linkage Models:** Five deterministic models and one probabilistic model were selected for evaluation based on review of the literature, evaluation of frequency distributions of citation variables, committee recommendations, current bibliography management tools, and the results of a pilot study. The model testing consisted of a record linkage of 10% samples of 1,180 WOS records to a dataset of 20,001 Medline records from which the controls had been withheld (Table 2), for a total of three trials for each of five record linkage models.

WOS DS 10% sample N= 1180		Linked to →	Medline DS N= 20,001
(Cases)	Original Journals n = 590		Original Journals n= 14149
(Controls)	Original Journals n = 590		(withheld, n= 0)
			Added Journals n = 5852

Table 2. Record Linkage Test Environment.

The deterministic record linkage modeling was performed using relational database queries. The probabilistic model was developed with LinkPlus, a record linkage program developed at the Centers for Disease Control and Prevention. LinkPlus computes probabilistic record linkage scores based on the theoretical framework developed by Fellegi and Sunter<sup>3</sup>, and computes M-Probabilities using the EM algorithm for maximum likelihood estimation.

**Deterministic Model #0 (DMatch0):** This model was evaluated to rule out the use of document titles as a single matching variable. Titles are nearly unique identifiers of documents as determined by frequency distributions, but due to inconsistencies in the recording of titles between Medline and WOS a model based on title as a single matching variable is not expected to perform well. The matching variable evaluated was Title (truncated at 50 characters).

**Deterministic Model #1 (DMatch1):** This model is based on the matchkey reported by Slach<sup>11</sup>. This

model was selected for evaluation because it was developed for use with bibliographic citation data, has simplicity, was not based on complex rules or weighting schemes, and had a low reported rate of false positives. The matching variables evaluated were Year, First Author Last Name (first 4 characters), and Begin page.

**Deterministic Model #2 (DMatch2):** This model was recommended by the Committee as being a standard for current good practice and is very similar to DMatch1 with the exception of using the full last name of the first author. The variables evaluated were Year, First Author Last Name, and Begin page.

**Deterministic Model #3 (DMatch3):** This model is based on the matching criteria used by the RefWorks bibliography management tool to identify duplicates. The variables evaluated were First Author Last Name, Year, and Title (truncated at 50 characters).

**Deterministic Model #4 (DMatch4):** This model was designed to avoid the use of author and title text strings which may be difficult to match because of variations in wording, spelling and punctuation. The variables evaluated were ISSN, Year, Volume, Issue, and Begin Page.

**Probabilistic Model (PScore):** There is an assumption of conditional independence in both the probabilistic scoring method and the EM algorithm (Winkler, 1999)<sup>4</sup>. The models assume that identifiers are independent, i.e. if there is a match on one variable there is not a second correlated variable that will have a very high probability of matching. For this reason both Journal ISSN and Journal Title elements were not combined in the list of candidate variables for probabilistic model development. Journal ISSN was selected over Journal Title because of less variability between the two databases. The final variables selected for use in the probabilistic linkage model and the matching parameters are shown in Figure 3.

Indirect method is employed, Field for Blocking: BeginPage						
Matching Parameters						
Matching Field	m-prob	u-prob	agree	disagree	matching method	
BeginPage	0.95000	0.00194	5.63580	-2.72506	exact	
Volume	0.95000	0.03916	2.90250	-2.69047	exact	
Year	0.95000	0.04783	2.72045	-2.68222	exact	
Issue	0.95000	0.15732	1.63679	-2.57103	exact	
EndPageDigit	0.95000	0.09993	2.04986	-2.63100	exact	
LastName	0.95000	0.00051	6.84846	-2.72637	generic string	
FirstInitial	0.95000	0.05884	2.53195	-2.67163	exact	
Title50	0.95000	0.00005	8.93527	-2.72679	generic string	
TitleEnd	0.95000	0.00339	5.12913	-2.72374	exact	

Figure 3. Probabilistic Record Linkage Matching Parameters

## Results

Model performance was stable across samples, and in the interest of saving space, the combined linkage results of the trials are summarized in Table 3. All of the deterministic models returned excess links due to false positive matches resulting from non-unique match keys. Two similar deterministic models based on name, year and page (DMatch1 & 2) performed well overall, but have problems with false negatives when there are differences in spelling and

punctuation of last names between datasets. The probabilistic model linkage obtained the highest AROC, and was significantly different from DMatch1 & 2 ( $P < .01$ ). The effect of fusion (the insertion of abstracts and MeSH terms from Medline into the WOS file) on data quality is summarized in Table 4. The increase in terms, nodes, and links in the visualization of a dataset prepared with probabilistic record linkage and information fusion methodology is summarized in Table 5.

Linkage Models	Confusion Matrix			Sensitivity	Specificity	AROC's	Model Weakness
<b>DMatch0</b>	N=3594	Matches	Non-Matches	.668	.958	.79 - .83	- Differences in wording and punctuation of titles between datasets
	Linked	1181	77				
	Unlinked	588	1748				
<b>DMatch1</b>	N=3543	Matches	Non-Matches	.974	.995	.987	-Differences in spelling of last name between datasets. -Two authors with same last name publishing at same time will be linked incorrectly – e.g., C. Friedman or last names beginning with “Van “
	Linked	1724	8				
	Unlinked	46	1765				
<b>DMatch2</b>	N=3542	Matches	Non-Matches	.958	.996	.978	-Differences in spelling of last name between datasets. -Two authors with same last name publishing at same time will be linked incorrectly – e.g., C. Friedman or last names beginning with “Van “
	Linked	1695	7				
	Unlinked	75	1765				
<b>DMatch3</b>	N=3556	Matches	Non-Matches	.649	.986	.80 - .83	-Differences in wording, spelling, or punctuation of Name or Title.
	Linked	1149	25				
	Unlinked	621	1761				
<b>DMatch4</b>	N=3542	Matches	Non-Matches	.879	.998	.93-.94	-Differences in indexing of articles by journal ISSN. WOS indexes AMIA conference proceedings under JAMIA ISSN, Medline indexes under proceedings ISSN. -Different use of print vs. electronic ISSN. -Missing data in matching variables
	Linked	1556	3				
	Unlinked	214	1769				
<b>PScore</b>	N=3540	Matches	Non-Matches	.995	.997	.999 – 1.0	
	Linked	1762	5				
	Unlinked	8	1765				

Table 3. Performance of Deterministic and Probabilistic Record Linkage Model

	% Complete Pre-Linkage	% Complete With Fusion
Abstract	.62	.80
KeyWords OR MeSH	.42	.97
Cited References	.94	.94
Records with Abstract AND Keywords AND Cited References	.40	.79

Table 4. Pre-Linkage and Post Probabilistic Linkage & Fusion Data Quality

	Pre-Linkage (Figure #)	With Fusion (Figure #)
Analysis Type	Document-Term Co-citation	Document-Term Co-citation
Publication Years	1976-1990	1976-1990
Thresholding (c/cc/ccv)	4/2/20	4/2/20
Burst Terms In Range	1,632	3,032
Nodes & Links	100 & 330	236 & 3,817

Table 5. Pre-Linkage and Post Probabilistic Linkage & Fusion Visualization Metrics

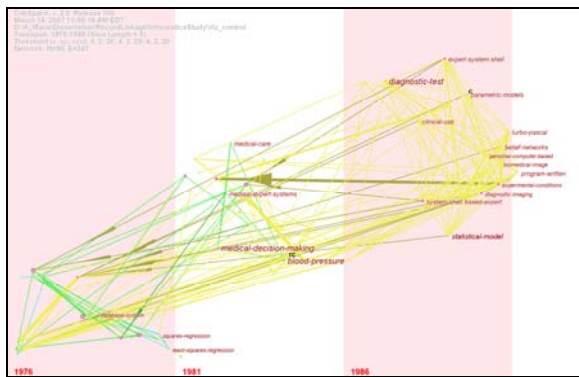


Figure 4. Terms available in visualization of WOS data, pre-linkage.

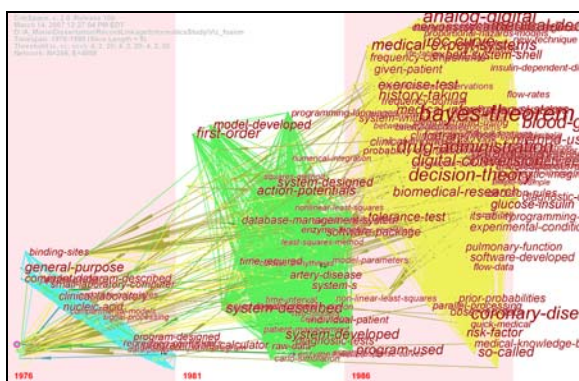


Figure 5. Terms available in visualization of WOS data, with record linkage and information fusion.

The effect of data preparation with record linkage and information fusion can readily be seen in a comparison of pre- and post-linkage visualizations (Figures 4 & 5). These visualizations have been generated using identical thresholds, and are presented without added pruning to demonstrate the added density of content that could be available to a user.

## Conclusion

Knowledge domain visualization has the potential for use by a wide range of users, notably scientists, clinicians, science policy researchers, and medical librarians. However if there is a deficiency or anomaly in the data used to generate visualizations, knowledge domain visualization also has the potential to mis-inform. Data preparation with record linkage and information fusion is a valid approach to addressing data quality and further enriching visualizations. While both deterministic and probabilistic models performed well in this evaluation, the probabilistic approach performed best by AROC measures. The deterministic models have

the weakness of an underlying assumption that there will not be multiple citations with the same author last name and same page occurring in the same year. This assumption will not hold true in all domains. While the context of this study was knowledge domain visualization, probabilistic record linkage also has the potential for practical applications in biomedical library use. This methodology could be applied to merging of multi-database searches, allow querying by MeSH terms with results ranked by citations counts, and be used for better management of bibliographies. Work is in progress to evaluate the application of record linkage methodology in additional knowledge domains and biomedical databases.

## References

1. Han J, Kamber J. Data Mining: Concepts and techniques. San Francisco:Morgan Kaufmann Publishers; 2001.
2. Newcombe HB, Kennedy JM, Axford AP. Automatic linkage of vital records. *Science*. 1959;130(3381): 954-959.
3. Fellegi IP, Sunter AB. A theory for record linkage. *Journal of the American Statistical Association*. 1969; 64(328): 1183-1210.
4. Winkler WE. The state of record linkage and current research problems. 1999. From <http://www.census.gov/srd/papers/pdf/rr99-04.pdf>.
5. Winkler WE. Data cleaning methods. Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation, Washington, DC. 2003.
6. Torra, V, Domingo-Ferrer J. Record linkage methods for multidatabase data mining. *Information fusion in data mining* (pp. 101-132). Berlin: New York: Springer; 2003.
7. Synnestevedt MB, Chen C, Holmes JH. CiteSpace II: Visualization and knowledge discovery in bibliographic databases. *AMIA '05*, Washington, DC. 2005:724-728.
8. Synnestevedt MB, Chen C, Holmes JH. Visual exploration of landmarks and trends in the medical informatics literature. *AMIA '05*, Washington, DC. 2005:1129.
9. Chen C. Searching for intellectual turning points: progressive knowledge domain visualization. *Proc Natl Acad Sci U S A*. 2004 Apr 6;101 Suppl 1:5303-10.
10. Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Jasist*. 2006;57(3):359-377.
11. Slach JE. Detection and elimination of duplicates from multidatabase searches. *Bulletin of the Medical Library Association*. 1985;73(3), 235-237.