

## Record Linkage References [william.e.winkler@census.gov](mailto:william.e.winkler@census.gov) (2004February11)

- Alvey, W. and Jamerson, B. (eds.) (1997), *Record Linkage Techniques -- 1997* (Proceedings of An International Record Linkage Workshop and Exposition, March 20-21, 1997, in Arlington VA), also published by National Academy Press (1999) and available at <http://www.fcsm.gov> under methodology reports.
- Armstrong, J. A. (2000), "Weight Estimation for Large Scale Record Linkage Applications," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 1-10.
- Armstrong, J. A., Block, C. and Saleh, M. (1999), "Record Linkage for Electoral Administration," *Statistical Society of Canada, Proceedings of the Survey Methods Section*, 57-64.
- Armstrong, J. A., and Mayda, J. E. (1993), "Model-based Estimation of Record Linkage Error Rates," *Survey Methodology*, 19, 137-147.
- Baxter, R., Christen, P., and Churches, T. (2003), "A Comparison of Fast Blocking Methods for Record Linkage," *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, Washington, DC, August 2003.
- Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False- Match Rates in Record Linkage," *Journal of the American Statistical Association*, **90**, 694-707.
- Belin, T. R. (1993) "Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment," *Survey Methodology*, **19**, 13-29.
- Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.
- Bilenko, M. and Mooney, R. J. (2003a), "Adaptive Duplicate Detection Using Learnable String Similarity Metrics," *Proceedings of ACM Conference on Knowledge Discovery and Data Mining*, Washington, DC, August 2003.
- Bilenko, M. and Mooney, R. J. (2003b), "On Evaluation and Training-Set Construction for Duplicate Detection," *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, Washington DC, August 2003.
- Borkar, V., Deshmukh, K. and Sarawagi, S. (2001), "Automatic Segmentation of Text into Structured Records," Association of Computing Machinery SIGMOD '01.
- Borthwick, A. (2002), "MEDD 2.0," (Conference Presentation, New York, NY, USA, February 2002), Available at <http://www.choicemaker.com>.
- Chaudhuri, S., Gamjam, K., Ganti, V., and Motwani, R. (2003), "Robust and Efficient Match for On-Line Data Cleaning," ACM SIGMOD '03.
- Christen, P., Churches, T. and Zhu, J.X. (2002) "Probabilistic Name and Address Cleaning and Standardization," (The Australian Data Mining Workshop, November, 2002), available at <http://datamining.anu.edu.au/projects/linkage.html>.
- Churches, T., Christen, P., Lu, J. and Zhu, J. X. (2002), "Preparation of Name and Address Data for Record Linkage Using Hidden Markov Models," *BioMed Central Medical Informatics and Decision Making*, **2** (9), available at <http://www.biomedcentral.com/1472-6947/2/9/>.
- Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003a), "A Comparison of String Metrics for Matching Names and Addresses," *International Joint Conference on Artificial Intelligence, Proceedings of the Workshop on Information Integration on the Web*, Acapulco, Mexico, August 2003.
- Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003b), "A Comparison of String Distance Metrics for Name-Matching Tasks," *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, Washington DC, August 2003.
- Cohen, W. W. and Richman, J. (2002), "Learning to Match and Cluster Entity Names," ACM SIGKDD '02.
- Copas, J. R., and Hilton, F. J. (1990), "Record Linkage: Statistical Models for Matching Computer Records," *Journal of the Royal Statistical Society*, **A**, **153**, 287-320.
- Cozman, F. G., Cohen, I., Circio, M. C. (2003), "Semi-Supervised Learning of Mixture Models," in (T. Fawcett and N. Mishra, eds.) *Proceedings of the Twentieth International Conference on Machine Learning*, 99-106.
- DeGuire, Y. (1988), "Postal Address Analysis," *Survey Methodology*, **14**, 317-325.
- Della Pietra, S., Della Pietra, V., and Lafferty, J. (1997), "Inducing Features of Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, 380-393.
- Deming, W. E., and Gleser, G. J. (1959), "On the Problem of Matching Lists by Samples," *Journal of the American Statistical Association*, **54**, 403-415.

- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, **B**, **39**, 1-38.
- Denis, F., Laurent, A., Gilleron, R., and Tomasi, M. (2003), "Text Classification and Co-Training from Positive and Unlabeled Examples," Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, International Conference on Machine Learning.
- Dhillon, I. S., Mallela, S., and Kumar, R. (2003), "A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification," *Journal of Machine Learning Research*, **3**, 1265-1287.
- Do, H.-H. and Rahm, E. "COMA – A system for flexible combination of schema matching approaches," Very Large Data Bases '02.
- Elfekey, M., Vassilios, V., and Elmagarmid, A. "TAILOR: A Record Linkage Toolbox," IEEE International Conference on Data Engineering '02.
- Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64**, 1183-1210.
- Friedman, J. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, **29** (5), 1389-1432.
- Gill, L. (1999), "OX-LINK: The Oxford Medical Record Linkage System," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 15-33.
- Getoor, L., Friedman, N., Koller, D., and Taskar, B. (2003), "Learning Probabilistic Models for Link Structure," *Journal Machine Learning Research*, **3**, 679-707.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer: New York.
- Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **89**, 414-420.
- Jin, L., Li, C., and Mehrotra, S. (2003) Efficient Record Linkage in Large Data Sets, 8th International Conference on Database Systems for Advanced Applications (DASFAA 2003) 26 - 28 March, 2003, Kyoto, Japan, to appear.
- Kim, J. J. and W. E. Winkler (1995), "Masking Microdata Files" American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 114-119.
- Kim, J. J., and Winkler, W. E. (2001), "Multiplicative Noise for Masking Continuous Data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, to appear.
- Koller, D. and Pfeffer, A. (1998), "Probabilistic Frame-Based Systems," Proceedings of the Fifteenth National Conference on Artificial Intelligence.
- Lafferty, J., McCallum, A., and Pereira, F. (2001), "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in Proceedings of the International Conference on Machine Learning, 282-289.
- Lahiri, P. A. and Larsen, M. D. (2004) "Regression Analysis with Linked Data," *Journal of the American Statistical Association*, **81**, to appear.
- Larsen, M. (1999), "Multiple Imputation Analysis of Records Linked Using Mixture Models," *Statistical Society of Canada, Proceedings of the Survey Methods Section*, 65-71.
- Lahiri, P., and Larsen, M. D. (2000), "Model-Based Analysis of Records Linked Using Mixture Models," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 11-19.
- Larsen, M. D., and Rubin, D. B. (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, **79**, 32-41.
- Lu, Q. and Getoor, L. (2003), "Link-based Classification," in (T. Fawcett and N. Mishra, eds.) *Proceedings of the Twentieth International Conference on Machine Learning*, 496-503.
- McCallum, A., Nigam, K., and Unger, L. H. (2000), "Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching," in *Knowledge Discovery and Data Mining*, 169-178.
- McCallum, A. and Wellner, B. (2003), "Object Consolidation by Graph Partitioning with a Conditionally-Trained Distance Metric," *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, Washington DC, August 2003.
- Malin, B., Sweeney, L., and Newton, E. (2003), "Trail Re-Identification: Learning Who You Are From Where You Have Been," Workshop on Privacy in Data, Carnegie-Mellon University, March 2003.
- Meng, X., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: the SEM Algorithm," *Journal of the American Statistical Association*, **86**, 899-909.
- Meng, X., and Rubin, D. B. (1993), "Maximum Likelihood Via the ECM Algorithm: A General Framework," *Biometrika*, **80**, 267-278.
- Michalowski, M., Thakkar, S., and Knoblock, C. A. (2003), "Exploiting Secondary Sources for Object

- Consolidation," *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, Washington DC, August 2003.
- Navarro, G. (2001), "A Guided Tour of Approximate String Matching," *Association of Computing Machinery Computing Surveys*, **33**, 31-88.
- Neiling, M. and Jurk, S. (2003), "The Object Identification Framework," *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, Washington DC, August 2003.
- Neter, J., Maynes, E. S., and Ramanathan, R. (1965), "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, **60**, 1005-1027.
- Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.
- Newcombe, H. B., Kennedy, J. M. Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, **130**, 954-959.
- Newcombe, H.B. and Kennedy, J. M. (1962) "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information" *Communications of the Association for Computing Machinery*, **5**, 563-567.
- Pasula, H. and Russell, S. (2001), "Approximate Inference for First-Order Probabilistic Languages," *Proceedings of the International Joint Conference on Artificial Intelligence*.
- Pollock, J. and Zamora, A. (1984), "Automatic Spelling Correction in Scientific and Scholarly Text," *Communications of the ACM*, **27**, 358-368.
- Porter, E. H., and Winkler, W. E. (1999), "Approximate String Comparison and its Effect in an Advanced Record Linkage System," in Alvey and Jamerson (ed.) *Record Linkage Techniques - 1997*, 190-199, National Research Council, National Academy Press: Washington, D.C.
- Rahm, E. and Do, H.-H. (2000), "Data Cleaning: Problems and Current Approaches," *IEEE Bulletin on Data Engineering*, 23 (4), .
- Ristad, E. S., and Yianilos, P. (1998), "Learning String-Edit Distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 522-531.
- Russell, S. (2001), "Identity Uncertainty," *Proceedings of IFSA-01*.
- Sarawagi, S. and Bhamidipaty, A. (2002), "Interactive Deduplication Using Active Learning," *Very Large Data Bases '02*.
- Scheuren, F., and Winkler, W. E. (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, **19**, 39-58.
- Scheuren, F., and Winkler, W. E. (1997), "Regression analysis of data files that are computer matched, II," *Survey Methodology*, **23**, 157-165.
- Sekar, C. C., and Deming, W. E. (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association*, **44**, 101-115.
- Taskar, B., Abdeel, P., and Koller, D. (2002), "Discriminative Probabilistic Models for Relational Data," *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.
- Taskar, B., Segal, E., and Koller, D. (2001), Probabilistic Classification and Clustering in Relational Data," *Proceedings of the International Joint Conference on Artificial Intelligence*.
- Taskar, B., Wong, M. F., Abdeel, P. and Koller, D. (2003), "Link Prediction in Relational Data," *Neural Information Processing Systems*, to appear.
- Taskar, B., Wong, M. F., and Koller, D. (2003), "Learning on Test Data: Leveraging "Unseen" Features," *Proceedings of the Twentieth International Conference on Machine Learning*, 744-751.
- Thibaudeau, Y. (1989), "Fitting Log-Linear Models When Some Dichotomous Variables are Unobservable," in *Proceedings of the Section on Statistical Computing, American Statistical Association*, pp. 283-288.
- Thibaudeau, Y. (1993), "The Discrimination Power of Dependency Structures in Record Linkage," *Survey Methodology*, **19**, 31-38.
- Titterton, D. M., Smith, A. F. M. and Makov, U. E. (1988), *Statistical Analysis of Finite Mixture Distributions*, New York: J. Wiley.
- Torra, V. (2000), "Re-Identifying Individuals Using OWA Operators," *Proceedings of the Sixth Conference on Soft Computing*, Iizuka, Fukuoka, Japan.
- Torra, V. (2003), "OWA Operators in Data Mining: Modeling and Re-Identification," preprint Feb. 2003.
- Vapnik, V. (2000), *The Nature of Statistical Learning Theory (2<sup>nd</sup> Edition)*, Berlin: Springer-Verlag.
- Wei, J. (2004), Markov Edit Distance, *IEEE PAMI*, 26 (3), 311-321.
- Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.

- Winkler, W. E. (1989a), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Fifth Census Bureau Annual Research Conference*, 145-155.
- Winkler, W. E. (1989b), "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage," *Survey Methodology*, **15**, 101-117.
- Winkler, W. E. (1989c), "Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 778-783.
- Winkler, W. E. (1990a), "Documentation of record-linkage software," unpublished report, Washington DC: Statistical Research Division, U.S. Bureau of the Census.
- Winkler, W. E. (1990b), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Assn.*, 354-359.
- Winkler, W. E. (1990c), "On Dykstra's Iterative Fitting Procedure," *Annals of Probability*, **18**, 1410-1415.
- Winkler, W. E. (1993a) "Business Name Parsing and Standardization Software," unpublished report, Washington, DC: Statistical Research Division, U.S. Bureau of the Census.
- Winkler, W. E. (1993b), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-279.
- Winkler, W. E. (1994), "Advanced Methods for Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472 (longer version report 94/05 available at <http://www.census.gov/srd/www/byyear.html>).
- Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox *et al.* (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384 (also available at <http://www.fcsm.gov/working-papers/winkler.pdf>).
- Winkler, W. E. (1997). "Producing Public-Use Microdata That are Analytically Valid and Confidential," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 41-50.
- Winkler, W. E. (1998). "Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata," *Research in Official Statistics*, **1**, 87-104.
- Winkler, W. E. (1999a). "The State of Record Linkage and Current Research Problems," *Statistical Society of Canada, Proceedings of the Survey Methods Section*, 73-80 (longer version also available at <http://www.census.gov/srd/www/byyear.html>).
- Winkler, W. E. (1999b), "Issues with Linking Files and Performing Analyses on the Merged Files," *Proceedings of the Sections on Government Statistics and Social Statistics, American Statistical Association*, 262-265.
- Winkler, W. E. (1999c), "Record Linkage Software and Methods for Administrative Lists," Eurostat, *Proceedings of the Exchange of Technology and Know-How '99*, also available at <http://www.census.gov/srd/www/byyear.html>.
- Winkler, W. E. (2000a), "Machine Learning, Information Retrieval, and Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20-29. (also available at <http://www.niss.org/affiliates/dqworkshop/papers/winkler.pdf>).
- Winkler, W. E. (2001a), "The Quality of Very Large Databases," *Proceedings of Quality in Official Statistics '2001*, CD-ROM (also available at <http://www.census.gov/srd/www/byyear.html> as report rr01/04).
- Winkler, W. E. (2001b), "Record Linkage," in A. H. El-Shaarawi and W. W. Piegorsch (eds.) *Encyclopedia on Environmetrics*, New York: J. Wiley.
- Winkler, W. E. (2002), "Record Linkage and Bayesian Networks," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear (also at <http://www.census.gov/srd/www/byyear.html>).
- Winkler, W. E. (2003a), "Methods for Evaluating and Creating Data Quality," *Proceedings of the IEEE Workshop on Cooperative Information Systems*, Sienna, Italy, January 2003.
- Winkler, W. E. (2003b), "Data Cleaning Methods," *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, Washington, DC, August 2003.
- Winkler, W. E. and Scheuren, F. (1991), "How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis," U.S. Bureau of the Census, Statistical Research Division Technical Report.
- Winkler, W. E. and Scheuren, F. (1995), "Linking Data to Create Information," *Proceedings of Symposium 95, From Data to Information - Methods and Systems*, Statistics Canada, 29-37.
- Winkler, W. E. and Scheuren, F. (1996), "Recursive Analysis of Linked Data Files," *Proceedings of the 1996 Census Bureau Annual Research Conference*, 920-935.
- Winkler, W. E. and Thibaudeau, Y. (1991), "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census," U.S. Bureau of the Census, Statistical Research Division Technical report.
- Yancey, W.E. (2000), "Frequency-Dependent Probability Measures for Record Linkage," *Proceedings of the*

- Section on Survey Research Methods, American Statistical Association, 752-757* (also at <http://www.census.gov/srd/www/byyear.html>).
- Yancey, W.E. (2002), "Improving EM Parameter Estimates for Record Linkage Parameters," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear (also at <http://www.census.gov/srd/www/byyear.html>).
- Yancey, W.E. (2003), "An Adaptive String Comparator for Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear (also at <http://www.census.gov/srd/www/byyear.html>).
- Yancey, W.E., and Winkler, W. E. (2003), "BigMatch Software," computer system, documentation available at <http://www.census.gov/srd/www/byyear.html>
- Yancey, W.E., Winkler, W.E., and Creecy, R. H. (2002) "Disclosure Risk Assessment in Perturbative Microdata Protection," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, Springer: New York.