Record Linkage References william.e.winkler@census.gov (2013Oct15)

Abowd, J. and Vilhuber, L. (2004), "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers (with discussion)," *Journal of Business and Economic Statistics*, 23 (2), 133-165.

Agichstein, E., and Ganti, V. (2004), "Mining Reference Tables for Automatic Text Segmentation," *ACM Knowledge Discovery and Data Mining Conference 2004*, 20-29.

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://wwww.fcsm.gov under methodology reports.

Ananthakrishna, R., Chaudhuri, S., and Ganti, V. (2002), "Eliminating Fuzzy Duplicates in Data Warehouse," *Very Large Data Bases 2002, 586-597*.

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," S*tatistical 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.

Arasu, A., Re, C., and Suciu, D. (2009), "Large Scale Deduplication with Constraints using Dedupalog," *Proceedings of ICDE*, 52-63.

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.

Bayardo, R. J., Ma, Y., and Srikant, R. (2007), "Scaling Up All Pairs Similarity Search," *WWW 2007*, Banff, Alberta, Canada, May 2007.

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.

Benjelloun, O., Garcia-Molina, H., Su, Q., and Widom, J. (2005), "Swoosh: A Generic Approach to Entity Resolution," Stanford University technical report, March 2005.

Bentley, J.L., and Sedgewick, R.A. (1997), "Fast Algorithms for Searching and Sorting Strings," *Proceedings of the Eighth ACM-SIAM Symposium on Discrete Algorithms*, 360-369.

Bhattacharya, I., and Getoor, L. (2004), "Iterative Record Linkage for Cleaning and Integration" *ACM SIGMOD Workshop on Data Mining and Knowledge Discovery 2004*, Paris, France.

Bhattacharya, I., and Getoor, L. (2006), "A Latent Dirichlet Allocation Model for Entity Resolution" *Proceedings of the 6th SIAM Conference on Data Mining* (*SDM '06)47-58 – best paper.*

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Bilenko, M., Basu, S., and Sahami, M. (2005), "Adaptive Product Normalization: Using Online Learning for Record Linkage in Comparison Shopping," *IEEE ICDM 2005*, Houston, TX, November 2005.

Bilenko, M., Kamath, B., and Mooney, R. J. (2006), "Adaptive Blocking: Learning to Scale Up Record Linkage," Proceedings of the 6th IEEE International Conference on Data Mining.

Bilenko, M., and Mooney, R. J. (2003a), "Adaptive Duplicate Detection Using Learnable String Similarity Metrics," *Proceedings of ACM Conference on Knowledge Discovery and Data Mi*ning, Washington, DC, August 2003, 39-48.

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.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., and Fienberg, S. (2003), "Adaptive Name Matching in Information Integration," *IEEE Intelligent Systems*, 18 (50), 16-23.

Bilke, A., and Naumann, F. (2005), "Schema Matching Using Duplicates," *IEEE International Conference on Data Engineering*.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Borkar, V., Deshmukh, K., and Sarawagi, S. (2001), "Automatic Segmentation of Text into Structured Records," Association of Computing Machinery SIGMOD 2001, 175-186.

Borthwick, A. (2002), “MEDD 2.0,” (Conference Presentation, New York, NY, USA, February 2002), Available at http://www.choicemaker.com.
Broadbent, K., and Iwig, W. (1999), “Record Linkage at NASS using AutoMatch,” FCSM Research Conference, http://www.fcsm.gov/99papers/broadbent.pdf .
Chambers, R. (2009), “Regression Analysis of Probability-Linked Data,” Statisphere, Volume 4, http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm.
Chaudhuri, S., Gamjam, K., Ganti, V., and Motwani, R. (2003), “Robust and Efficient Match for On-Line Data Cleaning,” ACM SIGMOD ’03, 313-324, http://www.sigmod.org/sigmod/sigmod03/eproceedings/.
Chaudhuri, S., Ganti, V., and Motwani, R. (2005), “Robust Identification of Fuzzy Duplicates,” *IEEE International Conference on Data Engineering*.
Chipperfield, J. O., Bishop, G. R., and Campbell, P. (2011), Maximum Likelihood estimation for contingency tables and logistic regression with incorrectly linked data, *Survey Methodology*, 37 (1), 13-24.
Christen, P. (2005), “Probabilistic Data Generation for Deduplication and Data Linkage Systems,” *Proceedings of the Sixth International Conference on Intelligent Data Engineering and Automated Learning* (IDEAL ’05), http://datamining.anu.edu.au/publications/2005/ideal-2005.pdf .
Christen, P. (2007), “Towards Parameter-free Blocking for Scalable Record Linkage,” Joint Computer Science Technical Report Series, Australia National University, http://cs.anu.edu.au/techreports/2007/TR-CS-07-03.pdf .
Christen, P. (2012), *Data Matching : Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer: New York, NY.
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.
Christen, P. and Goiser, C. (2005), “Assessing Deduplication and Data Linkage Quality: What to Measure?” Australasian Data Mining Conference, paper available at http://datamining.anu.edu.au/projects/linkage-publications.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 Knowledge Discovery and Data Mining Conference* 2002, 475-480.
Cohen, W. W., and Sarawagi, S. (2004), “Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods,” *Proceedings of the ACM Knowledge Discovery and Data Mining Conference 2005*, 89-98.
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.
Cooper, W. S. and Maron, M. E. (1978) Foundations of Probabilistic and Utility-Theoretic Indexing, J. Assoc. Comp. Mach. 25 (1), 67-80.
Cozman, F. G., Cohen, I., and 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.
Culotta, A., and McCallum, A. (2005), “Joint Deduplication of Multiple Record Types in Relational Data,” *CIKM* 2005.
Dong, X., Halevy, A., and Madhavan, J. (2005), “Reference Reconciliation in Complex Information Spaces,” *Proceedings of the ACM SIGMOD Conference*.
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., (2002) "COMA – A system for flexible combination of schema matching approaches," Very Large Data Bases 20, 610-621.

Dong, X. L., Haley, A., and Yu, C. (2009), "Data integration with uncertainty," *VLDB Journal*.

Draisbach, U., Naumann, F., Szott, S., and Wonneberg, O. (2012), "Adaptive Windows for Duplicate Detection," *Proceedings of the ICDE*, Washington, DC, USA, March 2012.

Elfekey, M., Vassilios, V., and Elmagarmid, A. (2002), "TAILOR: A Record Linkage Toolbox," *IEEE International Conference on Data Engineering 2002*, 17-28.

Faloutsos, C., and Lin, K.-I. (1995), "FastMap: A Fast Algorithm for Indexing, Data-mining and Visualization of Traditional and Multimedia Datasets," *Proceedings of the ACM SIGMOD Conference* (San Jose, California), New York: ACM, 163-174.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183-1210.

Ferrante, A. and Boyd, J. (2012), A transparent and transportable methodology for evaluating record linkage software, *J. Biomed. Informatics*, 45, 165-172.

Ferragina, P., and Grossi, R. (1999), "The string B-tree: a new data structure for string search in external Memory and its applications," *Journal of the Association of Computing Machinery*, 46, 236-280.

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.

Goldstein, H., Harron, K., and Wade, A. (2011) The analysis of record-linked data using multiple imputation with data value priors, Statistics in Medicine, 31 (28), 3481-3493, doi: 10.1002/sim.5508.

Gómez-Bao, J.,, Larriba-Pey, J.-L., and Ribes-Puig, J., (2008), "Record Linkage Performance for Large Data Sets," Conference on Knowledge and Information Management, Hong Kong, China.

Gómez-Bao, J., Martinez, N., Escalé, F., Ribes-Puig, J., Muntés-Mulero, V., and Larriba-Pey, J.-L. (2008), "Memoization techniques to improve Entity Resolution performance," technical report, DAMA-UPC, Computer Architecture Dept., Campus Nord UPC, Barcelona, Catalonia, Spain.

Gravano, L., Ipeirotis, P. G., Jagadish, H. V., Koudas, N., Muthukrishnan, and Srivastava, D. (2001), "Approximate String Joins in a Database (Almost) for Free," *Proceedings of VLDB*, 491-500.

Guisado-Gomez, J. (2009), "Changing Levenshtein function by Jaro-Winkler function," technical report, DAMA-UPC, Computer Architecture Dept., Campus Nord UPC, Barcelona, Catalonia, Spain.

Guha, S., Koudas, N., Marathe, A., and Srivastava, D. (2004), "Merging the Results of Approximate Match *Operations," Proceedings of the 30th VLDB Conference*, 636-647.

Hall, P. A. V. and Dowling, G. R. (1980), "Approximate String Comparison," *Association of Computing Machinery, Computing Surveys*, 12, 381-402.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Hernandez, M. and Stolfo, S. (1995), "The Merge-Purge Problem for Large Databases," *Proceedings of ACM SIGMOD 1995*, 127-138.

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2007), *Data Quality and Record Linkage Techniques*, New York, N. Y.: Springer.

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2010), "Record Linkage," in (D. W. Scott, Y. Said, and E. Wegman, eds.) *Wiley Interdisciplinary Reviews: Computational Statistics*, New York, N. Y.: Wiley, 2 (5), September/October, 535-543 .

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2010), "Data Quality," in (D. W. Scott, Y. Said, and E. Wegman, eds.) *Wiley Interdisciplinary Reviews: Computational Statistics*, New York, N. Y.: Wiley, 3 (1), January/February, 12-21.

Hjaltson, G., and Samet, H. (2003), "Index-Driven Similarity Search in Metric Spaces," *ACM Transactions On Database Systems*, 28 (4), 517-580.

Ikeda, M. M., and Porter, E. H. (2008), "Additional Results of Nationwide Matching of 2000 Census Data," Statistical Research Division Report, http://www.census.gov/srd/papers/pdf/rrs2008-02.pdf .

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*, 84, 414-420.

Jin, L., Li, C., and Mehrotra, S. (2002), "Efficient String Similarity Joins in Large Data Sets," UCI technical Report, Feb. 2002, http://www.ics.uci.edu/~chenli/pub/strjoin.pdf .

Jin,L., Li, C. and Mehrotra, S. (2003), "Efficient Record Linkage in Large Data Sets," *Eighth International Conference for Database Systems for Advanced Applications (DASFAA 2003)*, 26-28 March, 2003, Kyoto, Japan, http://www.ics.uci.edu/~chenli/pub/dasfaa03.pdf .

Kang, J., Lee, D., and Mitra, P. (2006), "Identifying Value Mappings for Data Integration: An Unsupervised Approach," *The International Conference on Web Systems Engineering (WISE)*, Springer LNCS 3806.

Kawai, H., Garcia-Molina, H., Benjelloun, O., Menestrina, D., Whang, E., and Gong, H. (2006), "P-Swoosh: Parallel Algorithm for Generic Entity Resolution," Stanford University CS technical report.

Kenig, B. and Gal. A. (2009), "Efficient Entity Resolution with MFI Blocks," Proceedings of VLDB, Lyon, France, August 2009.

Kenig, B. and Gal. A. (2013), MFI Blocks: An Effective Blocking Algorithm Entity Resolution, Information Systems, 38(6): 908-926 , http://dx.doi.org/10.1016/j.is.2012.11.008.

Kim, G. and Chambers, R. (2012a), Regression Analysis under Incomplete Linkage, *Computational Statistics and Data Analysis*, 56, 2756-2770.

Kim, G. and Chambers, R. (2012b), Regression Analysis under Probabilistic Multi-linkage, Statistica Neerlandica, 66 (1), 64-79.

Kim, H.-S., and Lee, D. (2007), "Parallel Linkage," CIKM '07.

Kim, H.-S., and Lee, D. (2010), "HARRA: Fast Iterative Hashed Record Linkage for Large-Scale Data Collections," EBDT '10, Lausanne, Switzerland, March 2010.

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*, CD-ROM.

Koepcke, H. and Rahm, E., (2010), "Frameworks for entity matching: a comparison," *Data and Knowledge Engineering*, 96 (2).

Koepcke, H., Thor, A., and Rahm, E., (2010), "Evaluation of entity resolution approaches on real-world problems," Proceedings of the VLDB Endowment, 31 (1).

Kolb, L. and Rahm, E. (2013), "Parallel Entity Resolution with Dedoop," Datenbank-Spektrum, 13 (1), 23-32, http://dbs.uni-leipzig.de/file/parallel_er_with_dedoop.pdf.

Koller, D., and Pfeffer, A. (1998), "Probabilistic Frame-Based Systems," *Proceedings of the Fifteenth National Conference on Artificial Intelligence.*

Koudas, N., Marathe, A., and Srivastava, D. (2004), "Flexible String Matching Against Large Databases in Practice," *Operations," Proceedings of the 30th VLDB Conference*, 1078-1086.

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., 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.

Lahiri, P. A., and Larsen, M. D. (2005) "Regression Analysis with Linked Data," *Journal of the American Statistical Association*, 100, 222-230.

Larsen, M. (1999), "Multiple Imputation Analysis of Records Linked Using Mixture Models," *Statistical Society of Canada Proceedings of the Survey Methods Section*, 65-71.

Larsen, M. D. (2005), "Hierarchical Bayesian Record Linkage Theory," Iowa State University, Statistics Department Technical Report.

Larsen, M. D., and Rubin, D. B. (2001), Alterative Automated Record Linkage Using Mixture Models, *Journal of the American Statistical Association*, 79, 32-41.

Liseo, B. and Tancredi, A. (2011), Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets, *Survey Methodology*, 27 (3), 491-505.

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.

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.

McCallum, A., Bellare, K., and Pereira, F. (2005), "A Conditional Random Field for Discriminatively-trained Finite-state String Edit Distance," *UAI* 2005.

McCallum, A., Nigam, K., and Unger, L. H. (2000, "Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching, in *Sixth ACM Conference of Knowledge Discovery and Data Mining*, 169-178, http://www.kamalnigam.com/papers/canopy-kdd00.pdf.

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.

McGovern, A., and Jensen, D. (2003), "Semi-Supervised Learning of Mixture Models," in (T. Fawcett and N. Mishra, eds.) *Proceedings of the Twentieth International Conference on Machine Learning*, 528-535.

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.

Michelson, M. and Knoblock, C. A. (2006), "Learning Blocking Schemes for Record Linkage," Proceedings of AAAI-2006.

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.

Michalowski, M., Thakkar, S., and Knoblock, C. (2004), "Exploiting Secondary Sources for Unsupervised Record Linkage," *Proceedings of the 30th VLDB Conference*, Toronto, Canada.

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.

Newcombe, H. B., and Smith, M. E. (1975), "Methods for Computer Linkage of Hospital Admission- Separation Records into Cumulative Health Histories," *Methods of Information in Medicine*, 14 (3), 118-125.

Newcombe, H. B., Smith, M. E., Howe, G. R., Mingay, G. R., Mingay, J., Strugnell, A., and Abbat, J.D. (1983), Reliability of Computerized Versus Manual Death Searches in a Study of the Health of Eldorado UraniumWorkers, *Computers in Biology and Medicine*, 13 (3), 157-169.

Norén, G. N., Orre, R., and Bate, A. (2005), "A Hit-Miss Model for Duplicate Detection in the WHO Drug Safety Database," Proceedings of the ACM KDD Conference.

Panel on Discriminant Analysis, Classification and Clustering, U.S. National Academy of Sciences, (1989), Discriminant Analysis and Clustering, Statistical Science, 4(1), 34-69.

Pasula, H., Baskara, M., Milch, B., Russell, S., and Shipster, I. (2003), "Identity Uncertainty and Citation Matching," *Neural Information Processing Systems 2003*.

Pasula, H., and Russell, S. (2001), "Approximate Inference for First-Order Probabilistic Languages," *Proceedings of the International Joint Conference on Artificial Intelligence*.

Pasula, H., Russell, S., Ostland, M., and Ritov, Y. (1999), "Tracking Many Objects with Many Sensors," *Proceedings of the Joint International 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, Washington, D.C: National Academy Press.

Rahm, E., and Do, H.-H. (2000), "Data Cleaning: Problems and Current Approaches," *IEEE Bulletin on Data Engineering*, 23 (4), 3-13.

Randall, S. M., Ferrante, A. M., Boyd, J. H., and Semmens, J. B. (2013), The effect of data cleaning on record linkage quality, BMC Medical Informatics and Decision Making, 13 (64), http://www.biomedicalcentral.com/1472-6947/13/64.

Rastogi, V., Dalvi, N., and Garofalakis, M. (2011), "Large-Scale Entity Resolution," *Proceedings of VLDB*, Seattle, WA, USA, September 2011.

Ravikumar, P., and Cohen, W. W. (2004), "A Hierarchical Graphical Model for Record Linkage," *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, Banff, Calgary, CA, July 2004, http://www.cs.cmu.edu/~wcohen/postscript/uai-2004.pdf .

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*, http:/www.cs.berkeley.edu/~russell/papers/ifsai01-identity.ps .

Sarawagi, S., and Bhamidipaty, A. (2002), "Interactive Deduplication Using Active Learning, *Very Large Data Bases 2002*, 269-278.

Sarawagi, S., Chakrabarti, S., and Godbole, S. (2003), "Cross-Training: Learning Probabilistic Mappings between Topics, in (Getoor, L., ed.), *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *177-186.*

Scannapieco, M. (2003), "DAQUINCIS: Exchanging and Improving Data Quality in Cooperative Information Systems," Ph.D. thesis in Computer Engineering, University of Rome "La Sapienza."

Scheuren, F. (1980), "Methods of Estimation for the 1973 Exact Match Study," *Studies from Interagency Data Linkages*, (Report No. 101, U.S. Social Security Administration).

Scheuren, F.,and Winkler, W. E. (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, 19, 39-58, also at http://www.fcsm.gov/working-papers/scheuren_part1.pdf .

Scheuren, F.,and Winkler, W. E. (1997), "Regression analysis of data files that are computer matched, II," *Survey Methodology*, 23, 157-165, http://www.fcsm.gov/working-papers/scheuren_part2.pdf.

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.

Shen, W., Li, X., and Doan, A. (2005), "Constraint-Based Entity Matching," *Proc. AAAI*.

Smith, M. E., and Newcombe, H. B. (1975), "Methods of Computer Linkage for Hospital Admission-Separation Records into Cumulative Health Histories," *Meth. Inform. Medicine*, 14(3), 118-125.

Steel, P., and Konshnik, C. (1999), "Post-Matching Administrative Record Linkage Between Sole Proprietorship Tax Returns and the Standard Statistical Establishment List," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 179-189.

Talburt, J. R. (2011), Entity Resolution and Information Quality, Morgan Kaufmann, Burlington, MA.

Tancredi, A., and Liseo, B. (2011), "A Hierarchical Bayesian Approach to Matching and Size Population Problems, *Ann. Appl. Stat.*, 5 (2B), 1553-1585.

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., Abbeel, P. and Koller, D. (2003), "Link Prediction in Relational Data," *Neural Information Processing Systems*, online at http://books.nips.cc/nips16.html .

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*, 283-288.

Thibaudeau, Y. (1993), "The Discrimination Power of Dependency Structures in Record Linkage," *Survey Methodology*, 19, 31-38.

Titterington, 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. (2004), "OWA Operators in Data Modeling and Re-Identification," *IEEE Transactions on Fuzzy Systems*, 12 (5) 652-660.

Vapnik, V. (2000), *The Nature of Statistical Learning Theory (2nd Edition)*, Berlin: Springer.

Wang, S., Schuurmans, D., Peng, F., and Zhao, Y. (2003), “Learning Mixture Models with the Latent Maximum Entropy Principal,” in (T. Fawcett and N. Mishra, eds.) *Proceedings of the Twentieth International Conference on Machine Learning*, 776-783 (also version for *IEEE Transactions on Neural Networks in 2004*).

Wei, J. (2004), Markov Edit Distance, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 (3), 311-321.

Weis, M. and Naumann, F. (2008), “Industry-Scale Duplicate Detection,” *Proceedings of PVLDB*, Auckland, New Zealand.

Whang, S. and Garcia-Molina, H.. (2012), “Joint Entity Resolution,” *Proceedings of the ICDE*, Washington, DC, USA, March 2012.

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, also at http://www.census.gov/srd/papers/pdf/rr2000-05.pdf .

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 Association*, 354-359 (available at *www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf*).

Winkler, W. E. (1990c), “On Dykstra’s Iterative Fitting Procedure,” *Annals of Probability*, 18, 1410-1415.

Winkler, W. E. (1991), “Error Model for Computer Linked Files,” P*roceedings of the Section on Survey Research Methods*, *American Statistical Association*, 472-477.

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," P*roceedings of the Section on Survey Research Methods*, *American Statistical Association*, 274-279, also http://www.census.gov/srd/papers/pdf/rr93-12.pdf .

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, D. A. Binder, B. N. Chinnappa, A. Christianson, Colledge, M. A., and P. S. Kott (eds.) *Business Survey Methods*, New York: J. Wiley, 355-384 (also available at http://www.fcsm.gov/working-papers/wwinkler.pdf).

Winkler, W. E. (1997). "Producing Public-Use Microdata That are Analytically Valid and Confidential," *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, 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,” *American Statistical Association*, *Proceedings of the Sections on Government Statistics and Social Statistics*, 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,” *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, 20-29. (also available at http://nisla05.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," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, CD-ROM (also at http://www.census.gov/srd/papers/pdf/rrs2002-05.pdf).

Winkler, W. E. (2003a), "Methods for Evaluating and Creating Data Quality," *Proceedings of the ICDT Workshop on Cooperative Information Systems*, Sienna, Italy, January 2003, longer version in *Information Systems* (2004), 29 (7), 531-550.

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. (2004a), "Re-identification Methods for Masked Microdata," in (J. Domingo-Ferrer and V. Torra, eds.) *Privacy in Statistical Databases 2004*, New York: Springer, 216-230, http://www.census.gov/srd/papers/pdf/rrs2004-03.pdf .

Winkler, W.E. (2004b), "Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems," in (J. Domingo-Ferrer and V. Torra, eds.) *Privacy in Statistical Databases 2004*, New York: Springer, 231-247, http://www.census.gov/srd/papers/pdf/rrs2004-06.pdf .

Winkler, W. E. (2004c), "Approximate String Comparator Search Strategies for Very Large Administrative Lists," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM (also report 2005/06 at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (2005a), "Record Linkage: Overview of Recent Developments and Applications," in (P. Falorsi, A. Pallara, A. Russo, eds.) *L'integrazione di dati di fonti diverse, Technice e applicaczioni del Record Linkage e metodi di stima basati sull'uso congiunto di fonti statistche e amministrative*, Rome: FrancoAngeli.

Winkler, W. E. (2006a), "Overview of Record Linkage and Current Research Directions," U.S. Bureau of the Census, Statistical Research Division Report http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf .

Winkler, W. E. (2006b), "Automatic Estimation of Record Linkage False Match Rates," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM, also at http://www.census.gov/srd/papers/pdf/rrs2007-05.pdf .

Winkler, W. E. (2008), "Data Quality in Data Warehouses," in (J. Wang, ed.) *Encyclopedia of Data Warehousing and Data Mining (2nd Edition)*.

Winkler, W. E. (2009), "Record Linkage," in (D. Pfeffermann and C. R. Rao, eds.) *Sample Surveys: Theory, Methods and Inference*, New York: North-Holland, 351-380.

Winkler, W. E. (2011), "Cleaning and using administrative lists: Enhanced practices and computational algorithms for record linkage and modeling/editing/imputation," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM.

Winkler, W. E. (2011), "Machine Learning and Record Linkage," *Proceedings of the International Statistical Institute*, Word Congress of Statistics, Dublin, Ireland, August 2011.

Winkler, W. E. (2011), "Cleaning and using administrative lists: Methods and fast computational algorithms for record linkage and modeling/editing/imputation," *Proceedings of the ESSnet Conference on Data Integration*, Madrid, Spain, November 2011 (http://www.ine.es/e/essnetdi_ws2011/ppts/Winkler.pdf).

Winkler, W. E. (2013a). "Record Linkage," in *Encyclopedia of Environmetrics*. J. Wiley.

Winkler, W. E. (2013b, Methods for adjusting statistical analyses for record linkage error, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM.

Winkler, W. E. (2013c), Matching and Record Linkage, submitted overview/survey article.

Winkler, W. E. (2014), Data Linkage: Overview, in (H. Goldstein, K. Harron, C. Dibben, eds.) *Methodological Developments in Data Linkage*, J. Wiley: New York, to appear.

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 91-9, http://www.census.gov/srd/papers/pdf/rr91-9.pdf.

Winkler, W. E., Yancey, W. E., and Porter, E. H. (2010), "Fast Record Linkage of Very Large Files in Support of Decennial and Administrative Records Projects," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM, also http://www.amstat.org/sections/srms/proceedings/y2010/Files/307067_57754.pdf .

Wright, J. (2010), "Linking Census Records to Death Registrations," Australia Bureau of Statistics Report 131.0.55.030.

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*, CD-ROM (also report RRS 2004/01at 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*, CD-ROM (also report RRS 2004/02 at http://www.census.gov/srd/www/byyear.html).

Yancey, W.E. (2005), "Evaluating String Comparator Performance for Record Linkage," research report RRS 2005/05 at http://www.census.gov/srd/www/byyear.html).

Yancey, W.E. (2007), "BigMatch: A Program for Extracting Probable Matches from Large Files," Statistical Division Research Report, http://www.census.gov/srd/papers/pdf/RRC2007-01.pdf .

Yancey, W.E. (2010), "Expected Number of Random Duplicates Within and Between Lists," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, CD-ROM, also http://www.amstat.org/sections/srms/proceedings/y2010/Files/307481_58657.pdf.

Yancey, W.E., and Winkler, W. E. (2004), "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*, New York: Springer, 135-152, http://www.census.gov/srd/papers/pdf/rrs2002-01.pdf .