

Data Wars Over Data Stores: challenges in medical data linkage

Denise de Vries
Anna Shillabeer
John F. Roddick



Barriers to eResearch

- Availability of data
 - Quantity
 - Quality
 - Applicability / suitability



Data Quality



- Why data is bad
- Re-using data for different purposes
- Consequences and implications of above

Dubious Data



- Poorly coded
- Incomplete
- Out-of-date
- Wrong data for tool being applied

Defence / Medical



Medical

- Different disciplines with different focus/roles/outcome measures
- Talk in acronyms
- Highly complex data
- Large quantities of pre-existing knowledge
- Highly distributed data sources
- Significant consequences for errors
- Politically sensitive

Defence

- Different disciplines with different focus/roles/outcome measures
- Talk in acronyms
- Highly complex data
- Large quantities of pre-existing knowledge
- Highly distributed data sources
- Significant consequences for errors
- Politically sensitive

Importantly - Similar techniques in terms of information management and analysis

Data Linkage



- Heterogeneous
- Non-standard
- Provenance
- Governance

Data Mining & Analysis



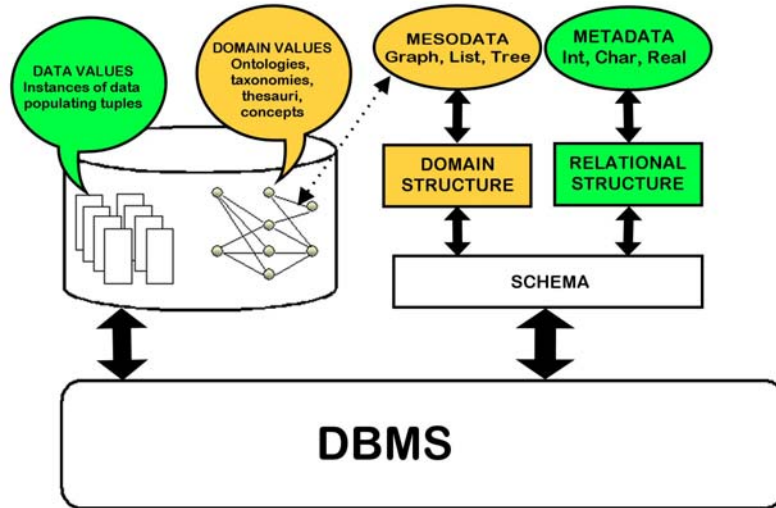
- Privacy
- Security
- Ethics

Semantic Considerations

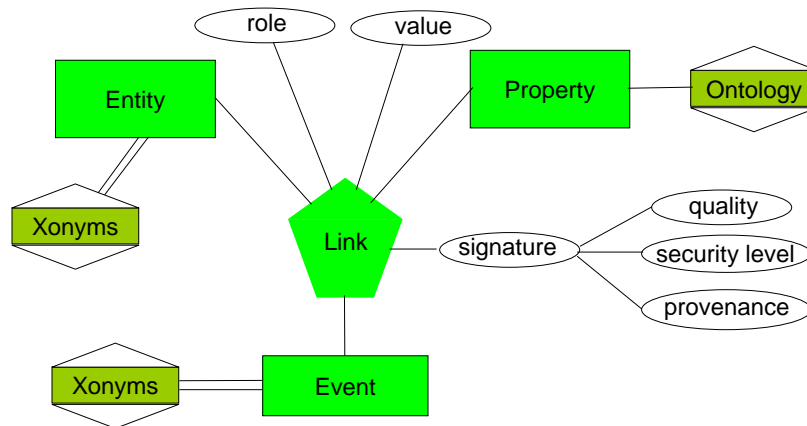


- Light Ontologies
- Concept Maps
- Taxonomies
- Medical Ontologies
 - SNOMED
 - ICD10 => ICD11
 - HL7, CPT-4, MeSH & LOINC
 - Different terminologies
 - Evolving terminologies

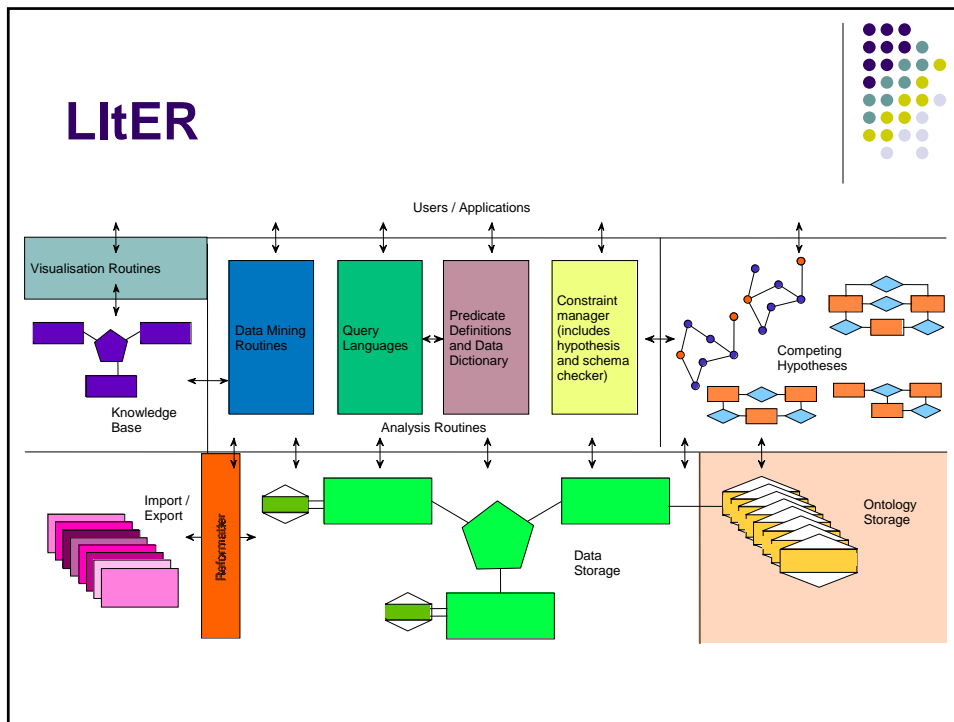
Mesodata



LitER Schema



LitER



Reporting Protocol

- Data testing
 - especially when linked
 - Is the data source valid?
- Results testing
 - replicated & validated
 - reported as non clinical test
- Methodology & statistical thresholds reported
- Data demographics reported
- Granularity check

Nurture Data for Future



- Complete data dictionaries
- Data secure against loss (local)
- Data secure against unauthorised access
- Communal repository ?
- Re-usable, sharable information
- Long term preservation of information

References



- Roddick, JF, Ceglar, A, De Vries, D & La-Ongsri, S 2008, 'Postponing Schema Definition : Low Instance-to-Entity Ratio (LIER) Modelling', in *Active Conceptual Modelling for Learning*, eds. PPS Chen & L Wong, Springer.
- Wahlstrom, K, Roddick, JF, Sarre, R, Estivill-Castro, V & de Vries, D 2008, 'Legal and Technical Issues of Privacy Preservation in Data Mining', in *Encyclopedia of Data Warehousing and Mining, 2nd edition*, ed. J Wang, IGI Publishing.
- de Vries, D & Roddick, JF 2007, 'The Case for Mesodata: An Empirical Investigation of an Evolving Database System', *Information and Software Technology*, vol. 49.
- Elmagarmid, AK, Ipeirotis, PG & Verykios, VS 2007, 'Duplicate Record Detection: A Survey', *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 1, 2007, pp. 1--16.
- Goertzen, R & Stausberg, J 2007, 'A grammar of integrity constraints in medical documentation systems', *Computer Methods and Programs in Biomedicine*, vol. 86, no. 1, 2007, pp. 93--102.
- Howe, HL, Lake, AJ & Shen, T 2007, 'Method to Assess Identifiability in Electronic Data Files', *American Journal of Epidemiology*, vol. 165, 2007, pp. 597--601.
- Miller, EA 2007, 'Solving the disjuncture between research and practice: Telehealth trends in the 21st century', *Health Policy*, vol. 82, no. 2, pp. 133-141.
- Patrick, J, Wang, Y & Budd, P 2007, 'An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology', paper presented at the Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007), Ballarat, Australia.
- Shillabeer, A 2007, 'An Automated Data Pattern Translation Process for Medical Data Mining', paper presented at the Medinfo 2007 Congress, Brisbane.
- Shillabeer, A & Pfitzner, D 2007, 'Determining Pattern Element Contribution in Medical Datasets', paper presented at the Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007), Ballarat, Australia.
- Ceglar, A, Morrall, R & Roddick, JF 2006, 'Mining Medical Administrative Data - The PKB System', paper presented at the ECML PKDD 2006 Workshop on Practical Data Mining: Applications, Experiences and Challenges, Berlin.

References



- de Vries, D 2006, 'Mesodata : Engineering Domains for Attribute Evolution and Data Integration', School of Informatics and Engineering, The Flinders University of South Australia.
- Karasti, H, Baker, K & Halkola, E 2006, 'Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network', *Computer Supported Cooperative Work (CSCW)*, vol. 15, pp. 321--358.
- Mooney, CH, De Vries, D & Roddick, JF 2006, 'A Multi-level Framework for the Analysis of Sequential Data', in *Data Mining: Theory, Methodology, Techniques, and Applications*, vol. 3755, eds. SJ Simoff & GJ Williams, Springer, Heidelberg, Germany, pp. 229-243.
- Patrick, J 2006, 'Metonymic and Holonymic roles and Emergent Properties in the SNOMED CT Ontology', paper presented at the Second Australasian Ontology Workshop (AOW 2006), Hobart, Australia.
- Rice, SP, Roddick, JF & de Vries, D 2006, 'Defining and Implementing Domains with Multiple Types using Mesodata Modelling Techniques', paper presented at the 3rd Asia-Pacific Conference on Conceptual Modelling, Hobart, Tasmania.
- Roddick, JF & de Vries, D 2006, 'Reduce, Reuse, Recycle: Practical Approaches to Schema Integration, Evolution and Versioning', paper presented at the 4th International Workshop on Evolution and Change in Data Management, Tucson, Arizona.
- Shillabeer, A & Roddick, JF 2006, 'Towards Role Based Hypothesis Evaluation for Health Data Mining', *Electronic Journal of Health Informatics*, vol. 1, no. 1, 2006, pp. 1-8.
- Shillabeer, A, Roddick, JF & DeVries, D 2006, 'On the Arguments Against the Application of Data Mining to Medical Data Analysis Editor: Peek, N.; Combi, C', paper presented at the Intelligent data Analysis in BioMedicine and Pharmacology (IDAMAP), Verona, Italy., August 25-26, 2006.
- Bass, J & Rosman, D 2005, 'What do you mean by 'Data Mining'?', paper presented at the ARC Health Data Mining Workshop, University of South Australia, Mawson Lakes, S.A., 11th April 2005.

References



- Bertino, E, Nai Fovino, I & Parasiliti Provenza, L 2005, 'A Framework for Evaluating Privacy Preserving Data Mining Algorithms', *Data Mining and Knowledge Discovery*, vol. 11, 2005, pp. 121--154
- Gorelick, MH, Alpern, ER, Singh, T, Snowdon, D, Holubkov, R, Dean, JM & Kuppermann, N 2005, 'Availability of Pediatric Emergency Visit Data from Existing Data Sources', *Academic Emergency Medicine*, vol. 12, no. 12, December 2005, pp. 1195-1200.
- Kelman, C 2005, 'Mining Linked Health data - a new frontier', paper presented at the ARC Health Data Mining Workshop, University of South Australia, Mawson Lakes, S.A., 11th April 2005.
- Mikkelsen, G & Aasly, J 2005, 'Consequences of impaired data quality on information retrieval in electronic patient records', *International Journal of Medical Informatics*, vol. 74, no. 5, pp. 387-394.
- Rice, SP, Roddick, JF & de Vries, D 2005, *Preventing Information Loss during Data Integration through Mesodata Modelling Techniques : Extended Report*, School of Informatics and Engineering, Flinders University.
- Shillabeer, A & Roddick, JF 2005, 'Reconceptualising Interestingness Metrics for Medical Data Mining', paper presented at the Australian Workshop on Health Data Mining, Mawson Lakes, SA.
- de Vries, D, Rice, S & Roddick, JF 2004, 'In Support of Mesodata in Database Management Systems', paper presented at the 15th International Conference on Database and Expert Systems Applications (DEXA 2004), Zaragoza, Spain.
- de Vries, D & Roddick, JF 2004, 'Facilitating Database Attribute Domain Evolution Using Mesodata', paper presented at the 3rd International Workshop on Evolution and Change in Data Management (ECDM2004), Shanghai.
- Fule, P & Roddick, JF 2004a, 'Detecting Privacy and Ethical Sensitivity in Data Mining Results', paper presented at the 27th Australasian Computer Science Conference (ACSC2004), Dunedin, New Zealand.
- Fule, P & Roddick, JF 2004b, 'Experiences in Building a Tool for Navigating Association Rule Result Sets', paper presented at the Australasian Workshop on Data Mining and Web Intelligence (DMWI2004), Dunedin, New Zealand.
- McGeehin, MA, Qualters, JR & Niskar, AS 2004, 'National environmental public health tracking program: bridging the information gap', *Environmental Health Perspectives*, vol. 112, no. 14, pp. 1409-13.

References



- Winkler, WE 2004, 'Methods for evaluating and creating data quality', *Information Systems*, vol. 29, no. 7, pp. 531-550.
- Augen, J 2003, 'Making Information Based Medicine Work', *Bio-IT World*, November 14th 2003.
- Berman, JJ 2003, 'Concept-match medical data scrubbing. How pathology text can be used in research', *Arch Pathol Lab Med*, vol. 127, no. 6, June 2003, pp. 680-686.
- Dimacs 2003, *DIMACS Working Group on Data Mining and Epidemiology*, Center for Discrete Mathematics & Theoretical Computer Science, Rutgers University.
- Kuonen, D 2003, 'Challenges in Bioinformatics for Statistical Data Miners', *Bulletin of the Swiss Statistical Society*, no. 46, October, pp. 10-17.
- Roddick, JF, Fule, P & Graco, WJ 2003, 'Exploratory Medical Knowledge Discovery : Experiences and Issues', *SigKDD Explorations*, vol. 5, no. 1, pp. 94-99.
- Roddick, JF, Hornsby, K & De Vries, D 2003, 'A Unifying Semantic Distance Model for Determining the Similarity of Attribute Values', paper presented at the 26th Australasian Computer Science Conference (ACSC2003), Adelaide, Australia.
- Wagner, MM, Robinson, JM, Tsui, FC, Espino, JU & Hogan, WR 2003, 'Design of a National Retail Monitor for Public Health Surveillance', *Journal of the American Medical Informatics Association.*, vol. 10, no. 5, September 2003, pp. 409-418.
- Cios, KJ & Moore, GW 2002, 'Uniqueness of medical data mining', *Artificial Intelligence in medicine*, no. 2002.
- McGee, MK 2002, *Early-Warning System could stem bioterrorist attacks and disease outbreaks*.
- Bresnahan, J 1997, 'Data mining - A delicate operation', *CIO Magazine*, June 15th.
- Dans, PE 1993, 'Looking for Answers in All the Wrong Places', *Annals of Internal Medicine*, vol. 119, no. 8, 15 October 1993, pp. 855-857.