REFLEXÕES SOBRE O ARTIGO: Overview and Framework for Data and Information Quality Research. Quadro Geral de Dados e Informações de Qualidade de Pesquisa.

STUART E. MADNICK and RICHARD Y. WANG. Massachusetts Institute of Technology; YANG W. LEE, Northeastern University and HONGWEI ZHU, Old Dominion University.

- O artigo apresenta uma visão geral da evolução e da paisagem atual de dados e pesquisa de qualidade da informação. Apresenta uma estrutura para caracterizar a pesquisa em duas dimensões: temas e métodos. Papéis representativos são citados para ilustrar os temas abordados e os métodos utilizados. É também identificado e discutido os desafios a serem abordados em pesquisas futuras.

- Questões que envolvem a qualidade dos dados e informações que fazem com que estas dificuldades variam em natureza da técnica (por exemplo, a integração de dados de fontes diferentes) para o não-técnica (por exemplo, a falta de uma estratégia coerente em toda a organização assegurar os interessados têm direito a informação correta no formato certo, no lugar certo e no tempo).

- Sobre a distinção entre a qualidade dos dados e informações de qualidade, há uma tendência para usar a qualidade dos dados para se referir a questões técnicas e de qualidade de informação para se referir a questões não técnicas.

- Em 1992, o MIT total de dados de Gestão da Qualidade (TDQM) programa foi oficialmente lançado para ressaltar a qualidade dos dados como uma área de investigação [Madnick e Wang 1992].

- Pesquisa de qualidade de dados no início, era focada principalmente no desenvolvimento de técnicas para consultar diversas fontes de dados e construção de grandes armazéns de dados.

- As primeiras pesquisas no programa desenvolvido no âmbito TDQM, que defendem a melhoria da qualidade de dados contínua, seguindo os ciclos de **Definir, Medir, Analisar, Melhorar** e [Madnick e Wang 1992]. O quadro se estende de Gestão quadro da Qualidade Total (TQM) para a melhoria da qualidade no domínio de fabricação [Deming 1982; Juran e Goferey 1999] para o domínio de dados. Pesquisas posteriores desenvolveram teorias, métodos e técnicas para os quatro ciclos de quadro TDQM.

- Do lado da indústria de campo da qualidade dos dados, os principais fabricantes de software começaram a implementar tecnologias de qualidade de dados em seus produtos e ofertas de serviços. No governo, a qualidade dos dados tornou-se um componente importante em muitos e-government e Enterprise Architecture (EA) iniciativas [OMB 2007]. No setor privado, as organizações adotaram variações na metodologia TDQM.

- Pesquisa de qualidade de dados, que começou há duas décadas, entrou em uma nova era, onde um número crescente de pesquisadores melhorar ativamente a compreensão

dos problemas de qualidade de dados e desenvolver soluções para problemas de qualidade de dados emergentes.

- UM QUADRO para caracterizar QUALIDADE DOS DADOS DE INVESTIGAÇÃO: O quadro era composto por sete elementos que a qualidade dos dados de impacto: (1) responsabilidades de gestão, (2) os custos de operação e segurança, (3) pesquisa e desenvolvimento; (4) produção; (5) distribuição; (6) gestão de pessoal e (7) A função legal. Pesquisa de qualidade de dados em 123 publicações até 1994 foram analisadas utilizando este framework. Embora este quadro fosse completo, faltava um conjunto de termos intuitivos para caracterizar a pesquisa de qualidade de dados e, portanto, não era fácil de usar. Além disso, os sete elementos não proporcionam granulosidade suficiente para fins de caracterização.

- Assim, um novo quadro tem duas dimensões: temas e métodos. É derivado de uma ideia simples: qualquer projeto de investigação de qualidade de dados aborda algumas questões (por exemplo, temas), utilizando certos métodos de investigação. Para cada dimensão, escolhe-se um pequeno conjunto de termos (isto é, palavras-chave) que têm significados intuitivos e deve abranger todas as características possíveis ao longo da dimensão. Essas palavras-chave são apresentados na Tabela I e suas explicações detalhadas são fornecidas nas próximas duas seções. Estes tópicos e método palavras-chave também são utilizadas para categorizar artigos submetidos para publicação no Journal of ACM dados e qualidade da informação.

- Para facilidade de uso, são escolhidas palavras-chave intuitivas e comumente utilizadas, tais como a mudança organizacional e integração de dados para a dimensão temas, e estudo de caso e econometria para a dimensão métodos. Os temas são agrupados em quatro categorias principais. Para os métodos de pesquisa, que estão listados em ordem alfabética, inclui-se um acordo com níveis de especificidade variáveis. Por exemplo, econometria é mais específico do que o método quantitativo. Este quadro dá aos usuários a flexibilidade de escolher um nível preferencial de especificidade na caracterização. Ao usar o framework para caracterizar uma determinada peça de investigação, o pesquisador escolhe um ou mais palavras-chave dimensão. Por exemplo, o papel "AIMQ: Uma Metodologia de Avaliação da Qualidade da Informação" [Lee et al. 2002] aborda a medição e avaliação do tópico e utiliza um método qualitativo particular (isto é, questionário), juntamente com um método quantitativo (ou seja, a análise estatística).

- Pode-se ver o quadro como uma matriz bidimensional, onde cada célula representa uma combinação tópico método. Pode-se colocar um trabalho de pesquisa em uma determinada célula de acordo com o tema abordado e do método utilizado. É possível colocar um papel em várias células se o documento aborda mais de um problema e / ou utiliza mais de um método. Um papel que utiliza um método mais específico pode também ser colocado na célula que corresponde a um modo mais geral. Obviamente, algumas células podem estar vazias ou pouco povoadas, tais como as células correspondentes a certas combinações de temas técnicos (por exemplo, integração de

dados) e métodos das ciências sociais (por exemplo, pesquisa-ação). Os pesquisadores são incentivados a considerar empregar mais de um método de pesquisa, incluindo um ou mais métodos quantitativos com um ou mais métodos qualitativos.

- Tópicos da Pesquisa: A qualidade dos dados é um campo multidisciplinar. Resultados de pesquisas existentes mostram que os pesquisadores estão principalmente a operar em duas disciplinas principais: sistemas de informações gerenciais (MIS) e Ciência da Computação (CS).

- Métodos da Pesquisa: Assim como existe uma infinidade de temas de pesquisa, há uma ampla gama de métodos de investigação adequados para a investigação de qualidade de dados. É identificado 14 categorias de alto nível de métodos de pesquisa.

Table I. Topics and Methods of Data Quality Research

Topics	Methods
1. Data quality impact	1. Action research
1.1 Application area (e.g., CRM, KM, SCM, ERP)	2. Artificial Intelligence
1.2 Performance, cost/benefit, operations	3. Case study
1.3 IT management	4. Data mining
1.4 Organizational change, processes	5. Design science
1.5 Strategy, policy	6. Econometrics
2. Database related technical solutions for data quality	7. Empirical
2.1 Data integration, data warehouse	8. Experimental
2.2 Enterprise architecture, conceptual modeling	9. Mathematical modeling
2.3 Entity resolution, record linkage, corporate householding	10. Qualitative
	11. Quantitative
2.4 Monitoring, cleansing	12. Statistical analysis
2.5 Lineage, provenance, source tagging	13. System design, implementation
2.6 Uncertainty (e.g., imprecise, fuzzy data)	14. Theory and formal proofs
3. Data quality in the context of computer science and IT	
3.1 Measurement, assessment	
3.2 Information systems	
3.3 Networks	
3.4 Privacy	
3.5 Protocols, standards	
3.6 Security	
4. Data quality in curation	
4.1 Curation - Standards and policies	
4.2 Curation - Technical solutions	

Fonte: S.E Madnick et al. 2009.

Conclusão: Investigação da qualidade dos dados tem feito progressos significativos nas últimas duas décadas. Desde o trabalho inicial realizado no programa TDQM (ver web.mit.edu / TDQM) e mais tarde o programa IQ (veja mitiq.mit.edu) no MIT, um número crescente de pesquisadores da ciência da computação, MIS, e outras disciplinas têm formaram uma comunidade que realiza ativamente pesquisa de qualidade de dados.

- Pesquisas são necessárias para o desenvolvimento de técnicas para a gestão e melhorar a qualidade dos dados nestas novas formas. Novas formas de entrega de informações também surgiram. Além da arquitetura cliente-servidor tradicional, a arquitetura orientada a serviços tem sido amplamente adotado como mais informação é agora

entregues através da Internet para os terminais tradicionais, bem como para dispositivos móveis.