Qualidade de Dados na Web

Integração de Dados e Warehousing

Caroline Pereira Medeiros

Roteiro

- Motivação
- Dados e Informação
- Qualidade
  - Qualidade de Dados
- Dados na Web
  - Qualidade de Dados na Web
- Conclusão
- Referências

Motivação

- Pesquisa de 2014
  - **1,7 megabytes/s** de novas informações digitais criadas em **2020**
  - **4.4 ZB** de dados hoje para **44ZB** até 2020
  - ↑ **10%** em **acessibilidade** dos dados => ↑ **US $ 65 milhões** na receita*
  - Menos de **0.5%** dos dados foram **analisados**

*Empresas da revista *Fortune 1000*

Pesquisa retirada do site *Forbes*

Motivação

- Pesquisa de 2015*
  - **52 %** dos cientistas de dados consideram tratar a **má qualidade de dados** como um dos **maiores desafios**
  - Quase **80%** consideram **escassa a quantidade de profissionais** de extração, limpeza e enriquecimento de dados

*150 cientistas participaram da pesquisa

Pesquisa retirada do site *Baseline*

Motivação

- Pesquisa de 2016*
  - **63% utilizam os dados** para realizar **operações** no dia-a-dia
  - **60% utilizam os dados** para melhor **entender** seus **clientes**
  - **51%** consideram a **qualidade dos dados** "moderada" ou "significantemente" **deficiente** na empresa (ou não conhecem a qualidade de seus dados)

*400 empresas e profissionais de TI participaram da pesquisa

Pesquisa retirada do site *Baseline*

Dados e Informação

- Dados
  - Representação do mundo real
  - Podem ou não ser utilizáveis
  - Representado por meio de entidades e atributos

- Informação
  - Após inserido um contexto o dado passa a ser informação
  - A qualidade da informação depende de sua utilidade
  - Sua qualidade está diretamente ligada à qualidade do dado

Qualidade

- O que é?
  - Capacidade de atingir o(s) efeito(s) pretendido(s)*
  - Concordância com os requisitos exigidos para o produto
  - Concordância com os requisitos exigidos pelo usuário
  - Termo de difícil definição e inserido em diferentes contextos

*Definição retirada do Google

Qualidade

- O que é preciso para medir qualidade?
  - É preciso definir o que é qualidade para cada aplicação
  - Fazer um estudo das propriedades que traduzem a aplicação
  - Traduzir estas propriedades em características mensuráveis
  - Estar atento as mudanças nas necessidades do consumidor

Qualidade de Dados

- Agentes:
  - Produtores/Publicadores
  - Administradores
  - Consumidores

Qualidade de Dados

- Alguns aspectos de QD:
  - Difícil avaliação pois os dados são multifacetados
  - Dimensões que dependem de:
    - Conjunto de dados
    - Das regras de negócios
    - Consumidor final
  - Dinamismo da avaliação de qualidade de dados
  - Reincidência de erros

Qualidade de Dados

- Categorias para QD:
  - Intrínseca
  - Acessibilidade
  - Contextual
  - Representacional

Qualidade de Dados

- **Intrínseca**
  - Precisão
  - Objetividade
  - Credibilidade
  - Reputação

Qualidade de Dados

- **Acessibilidade**
  - Acesso
  - Segurança

Qualidade de Dados

- **Contextual**
  - Relevância
  - Valor Agregado
  - Temporalidade
  - Completude
  - Quantidade de Dados

Qualidade de Dados

- **Representacional**
  - Interpretabilidade
  - Entendimento fácil
  - Representação concisa
  - Representação consistente

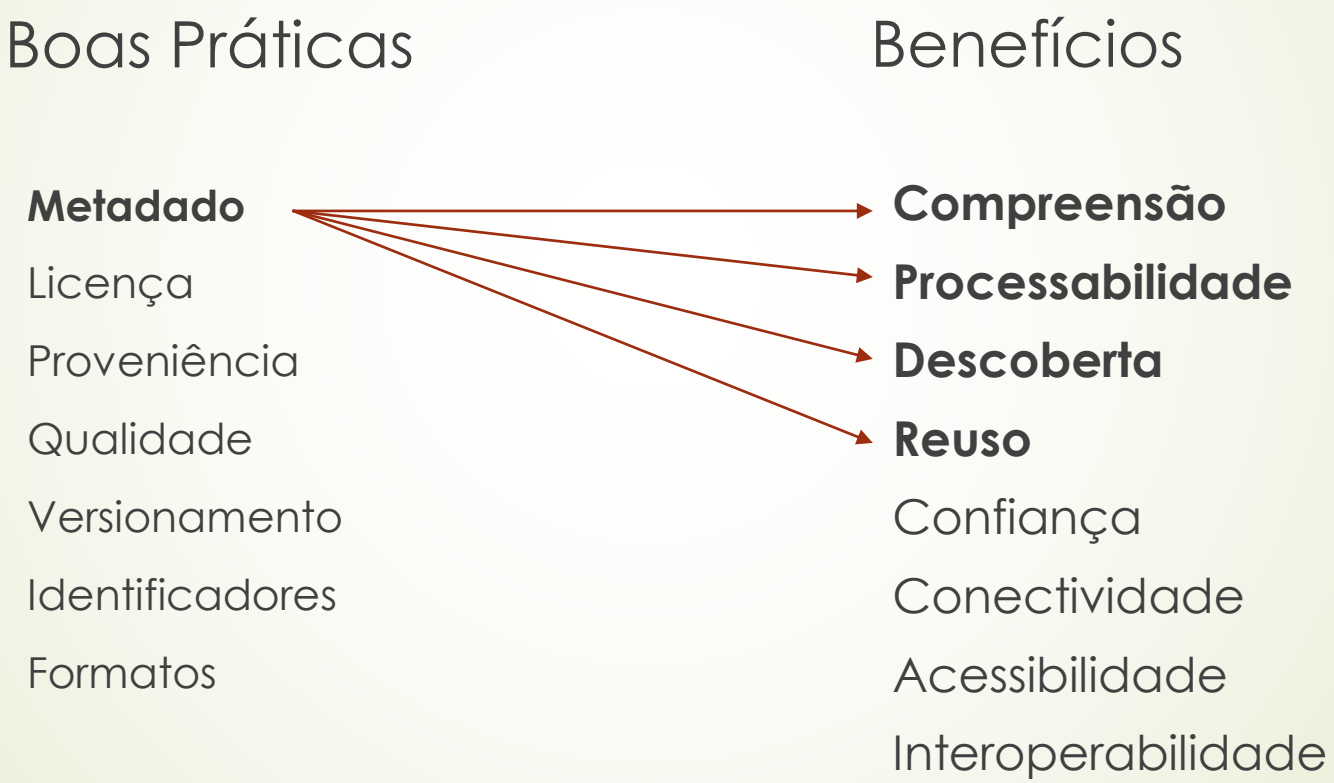
Dados na Web

- Características:
  - Grafo direcionado gigante
  - Web 1.0 (1989)
    - Links navegacionais
    - Links transacionais
  - Web 2.0 (2004)
    - Compartilhamento de conteúdo
    - Serviços na nuvem
    - Conexão direta entre pessoas
  - Web 3.0 (futuro)
    - Web Semântica
    - *Linked data*

Dados na Web

- Classificação de alguns problemas em dados Web:
  - Representacional
  - Conceitual
  - Temporal

Qualidade de Dados na Web



Qualidade de Dados na Web

Boas Práticas

Metadado

Licença

Proveniência

Qualidade

Versionamento

Identificadores

Formatos

Benefícios

Compreensão

Processabilidade

Descoberta

Reuso

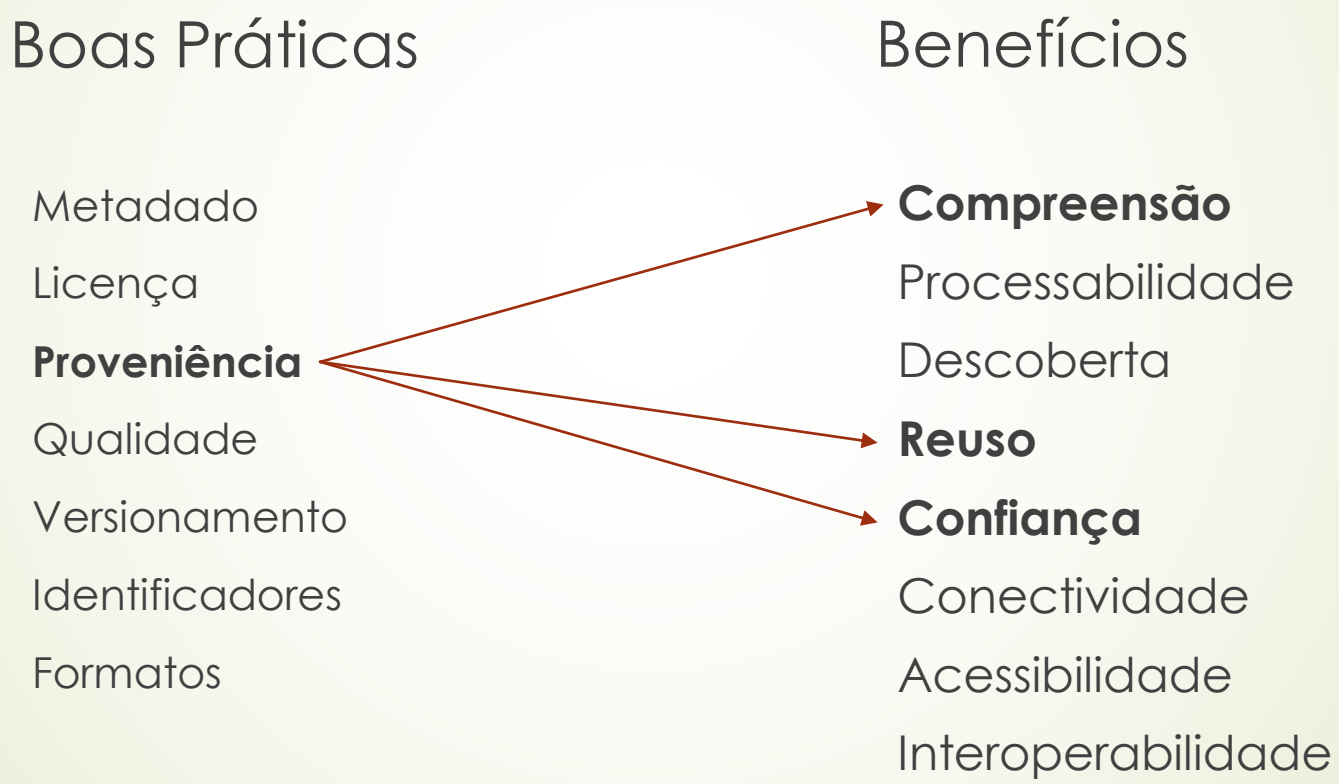
Confiança

Conectividade

Acessibilidade

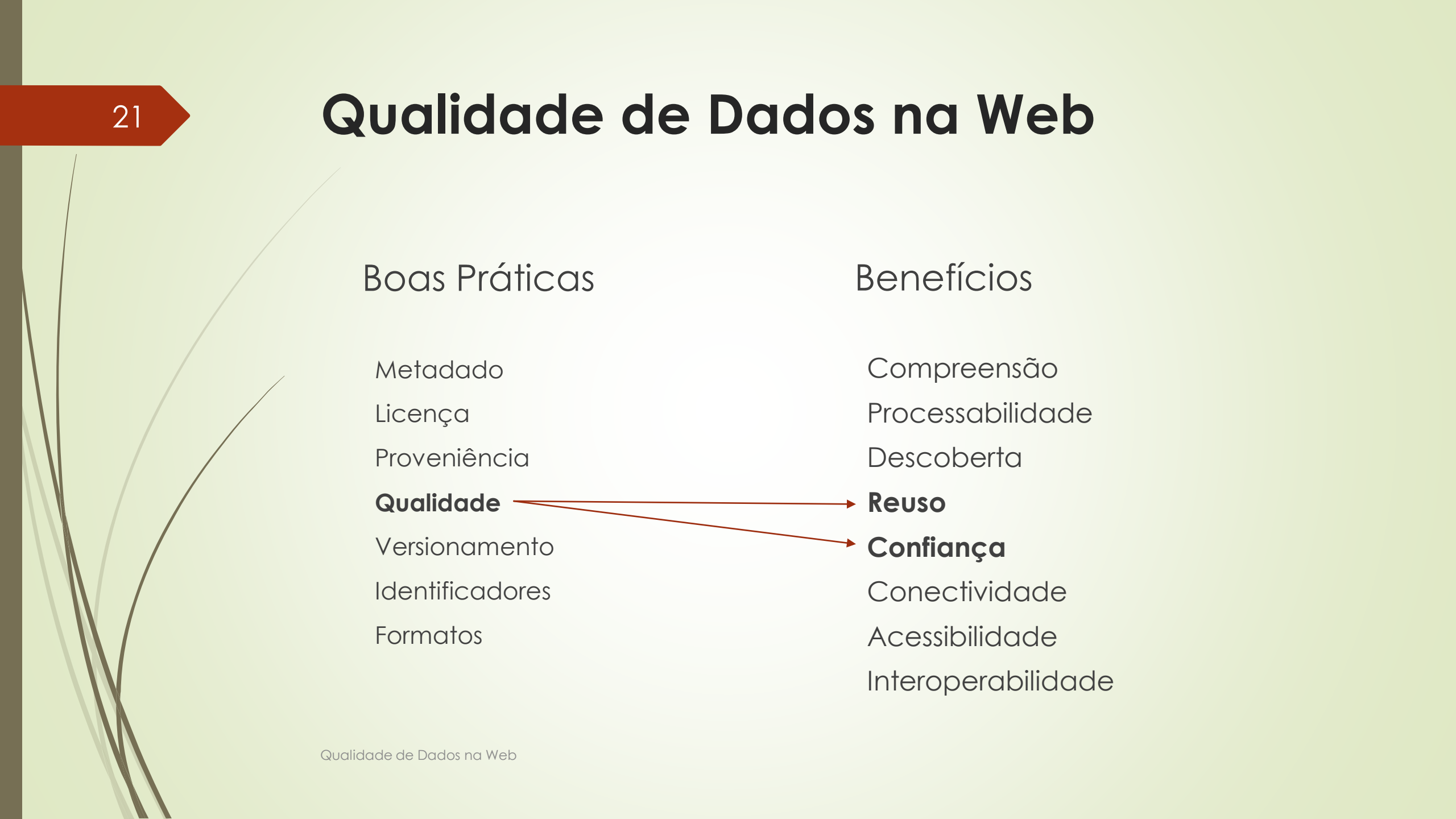
Interoperabilidade

Qualidade de Dados na Web



Qualidade de Dados na Web

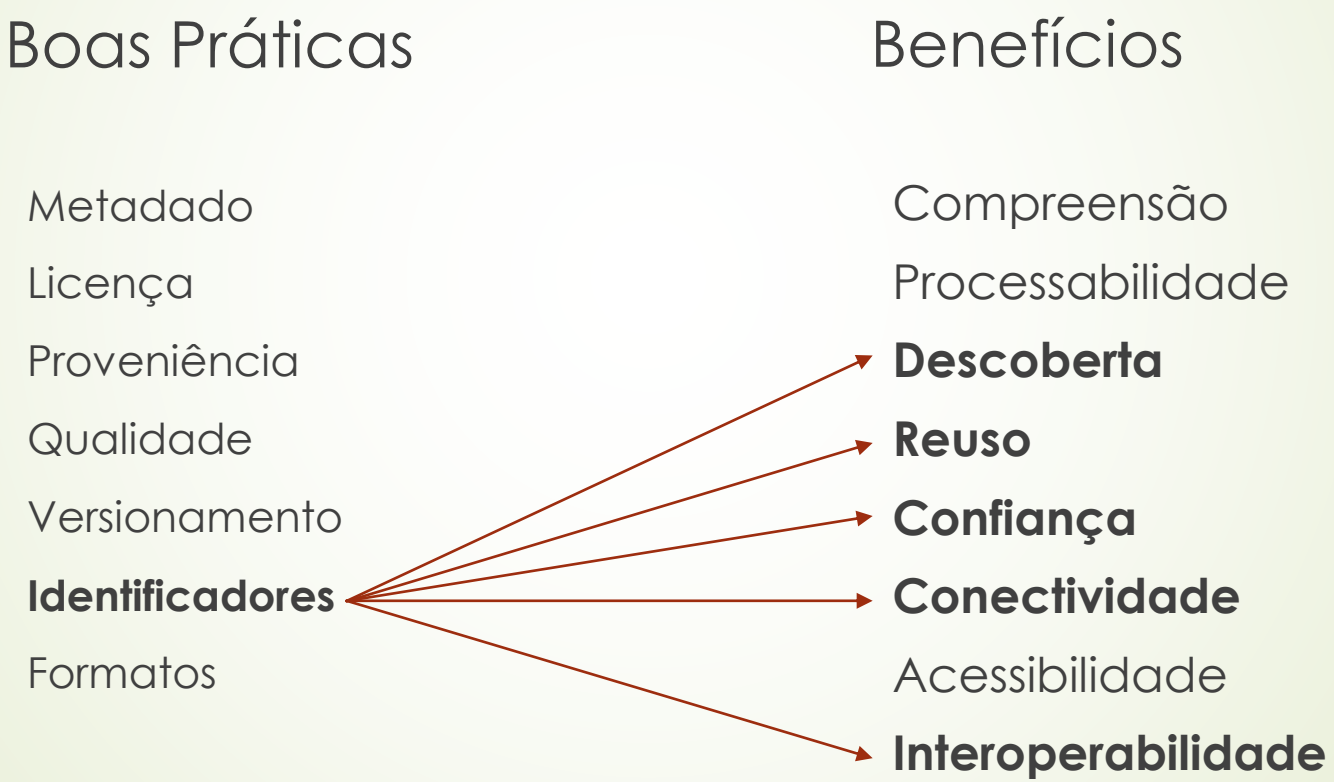
Boas Práticas	Benefícios
Metadado	Compreensão
Licença	Processabilidade
Proveniência	Descoberta
Qualidade	**Reuso**
Versionamento	**Confiança**
Identificadores	Conectividade
Formatos	Acessibilidade
	Interoperabilidade



Qualidade de Dados na Web

Boas Práticas	Benefícios
Metadado	Compreensão
Licença	Processabilidade
Proveniência	Descoberta
Qualidade	**Reuso**
Versionamento	**Confiança**
Identificadores	Conectividade
Formatos	Acessibilidade
	Interoperabilidade

Qualidade de Dados na Web



Qualidade de Dados na Web

Boas Práticas

Metadado

Licença

Proveniência

Qualidade

Versionamento

Identificadores

Formatos

Benefícios

Compreensão

Processabilidade

Descoberta

Reuso

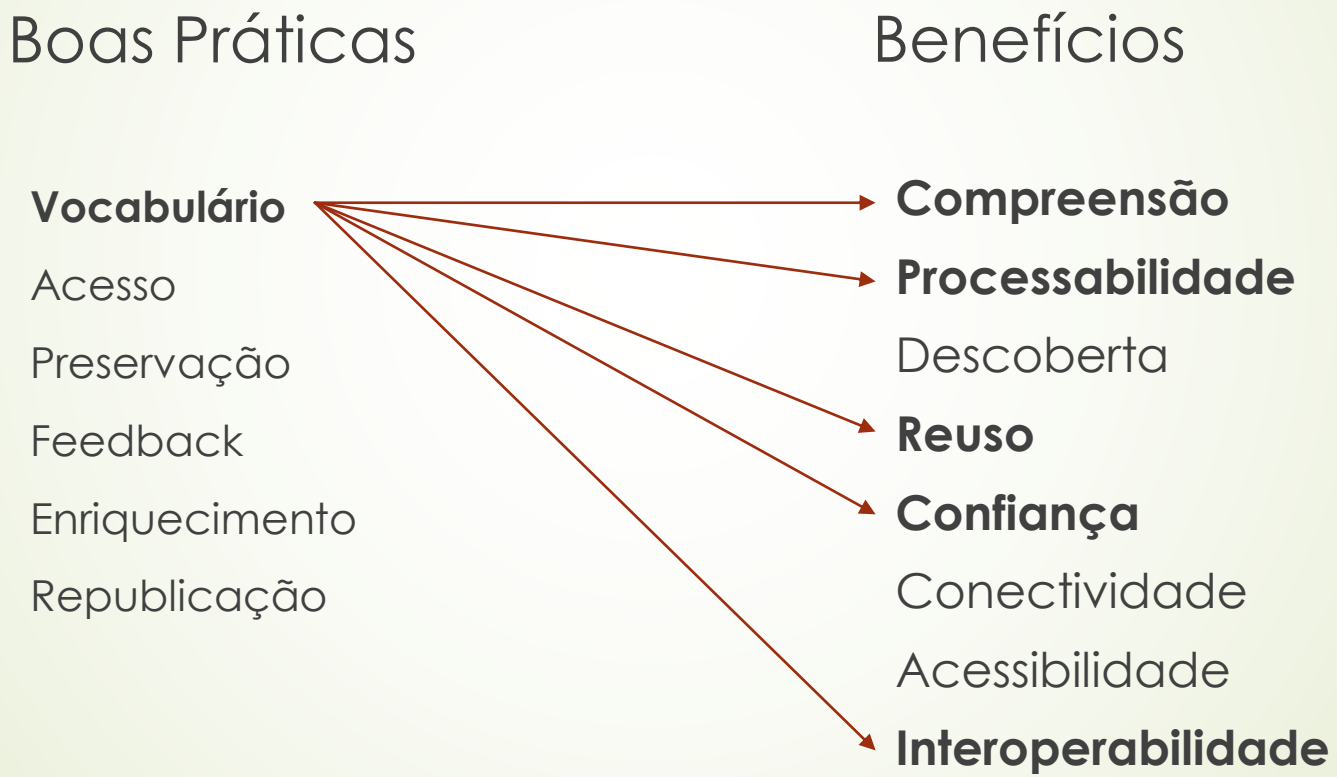
Confiança

Conectividade

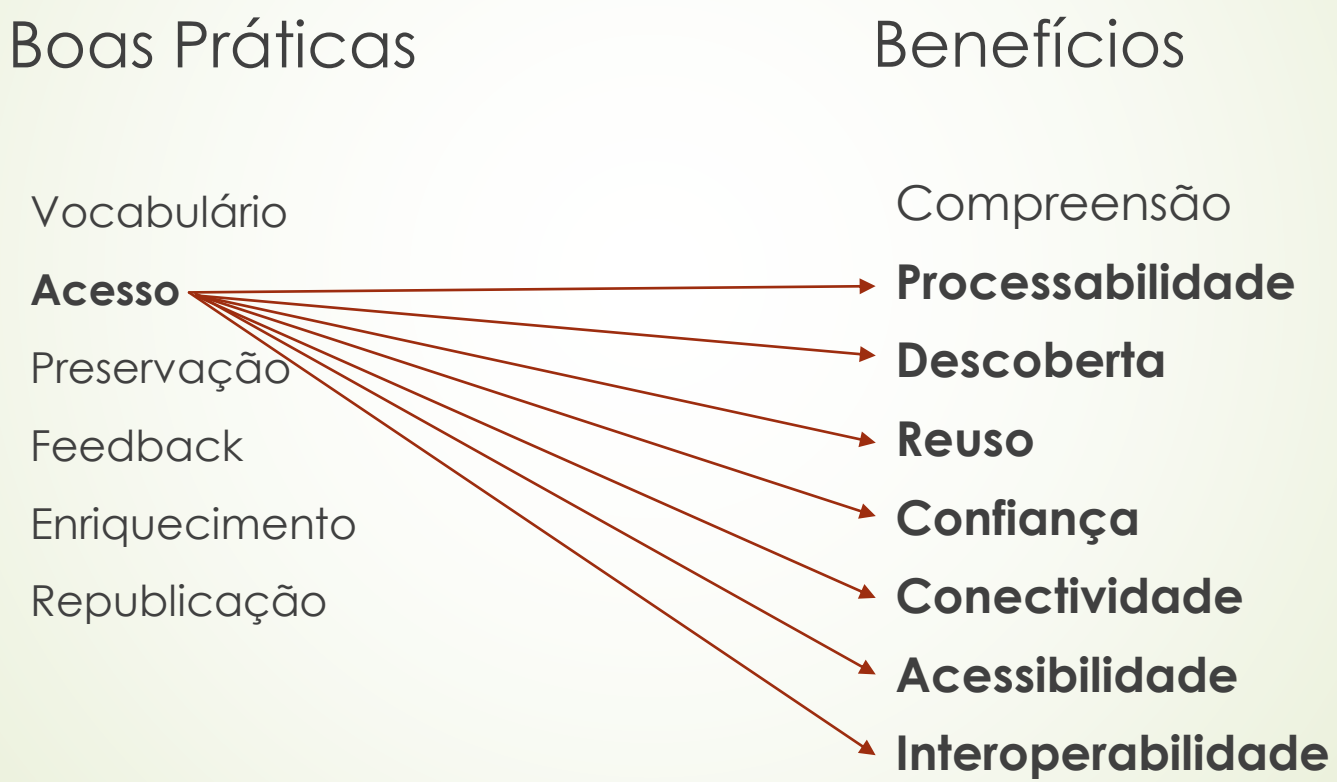
Acessibilidade

Interoperabilidade

Qualidade de Dados na Web



Qualidade de Dados na Web



Qualidade de Dados na Web

Boas Práticas

Vocabulário

Acesso

Preservação

Feedback

Enriquecimento

Republicação

Benefícios

Compreensão

Processabilidade

Descoberta

Reuso

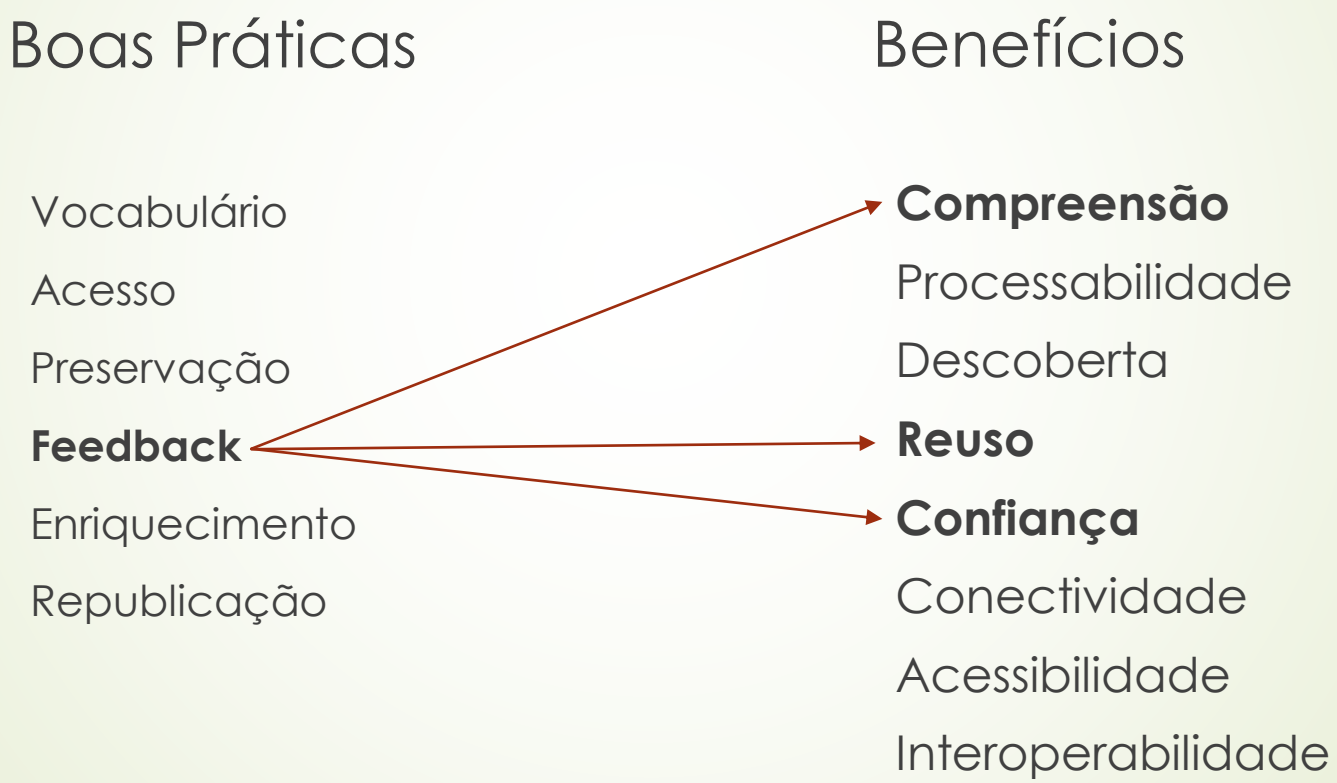
Confiança

Conectividade

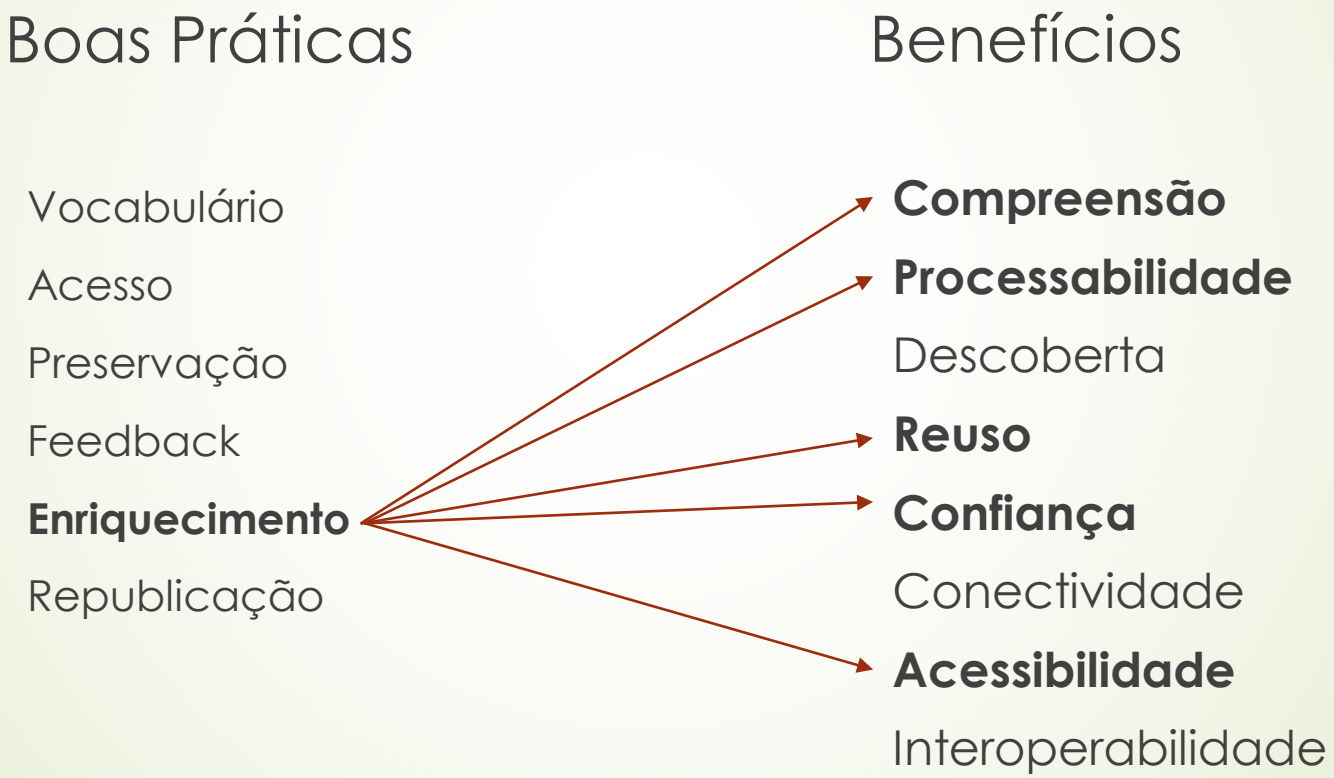
Acessibilidade

Interoperabilidade

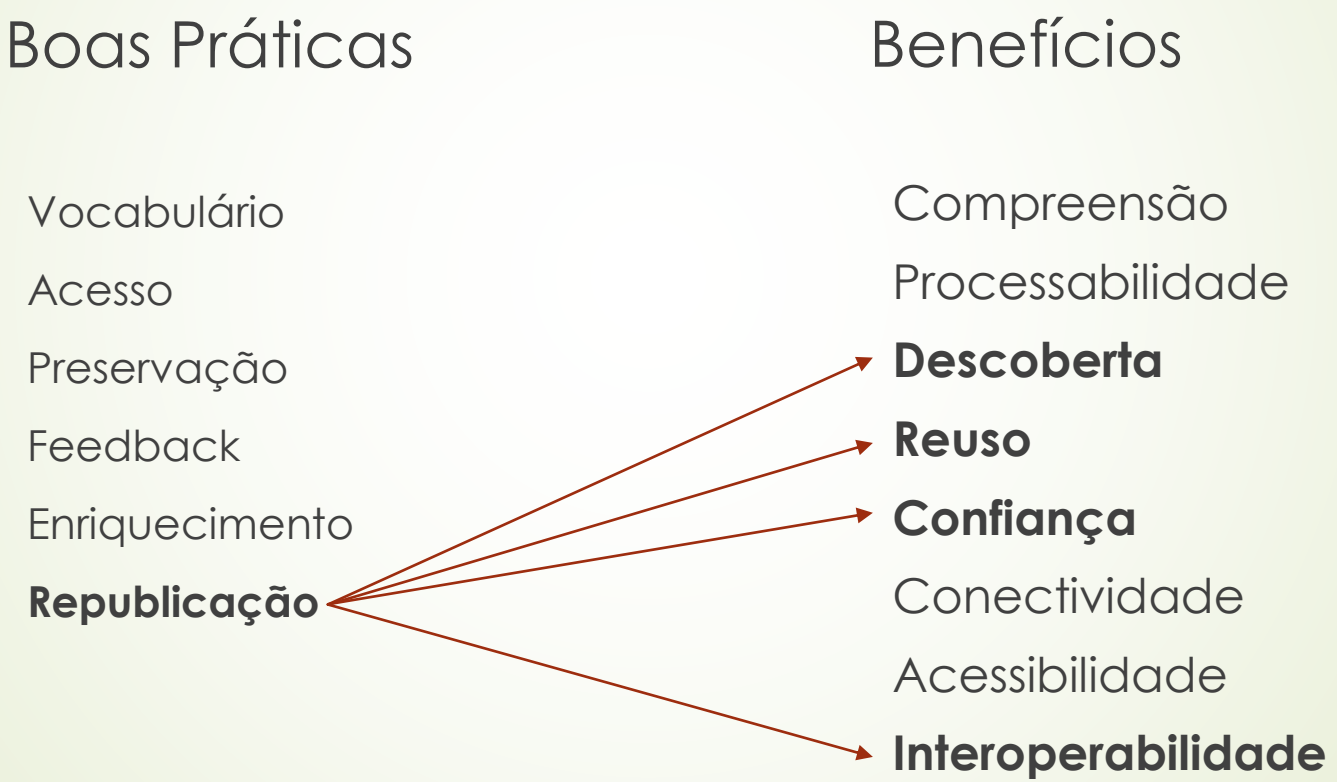
Qualidade de Dados na Web



Qualidade de Dados na Web



Qualidade de Dados na Web



Concluindo...

- Aumento expressivo na quantidade de dados disponíveis
- Novas formas de interação
  - Novas formas de gerar dados
  - Novos desafios
- Qualidade de dados com base nas necessidades do usuário
- Preparação para a chegada da Web Semântica e *Linked data*

Referências

- *Surprising Statistics About Big Data*
- *Big Data: 20 Mind-Boggling Facts Everyone Must Read*
- *Why Data Scientists Don't Have Time to Do Analysis*
- *What Companies Must Do to Get Value From Big Data*
- *What is Quality? Learn how each of eight well-known gurus answers this question*
- Diane M. Strong, Yang W. Lee, and Richard Y. Wang, 1997. "*Data Quality in Context*"
- Maria C. M. Batista, 2006. "*Information Quality Analysis in a Data Integration System*"
- *Data on the Web Best Practices*

Obrigada