Gerenciamento de Anotações de Biosseqüências utilizando Associação entre Ontologias e Esquemas XML

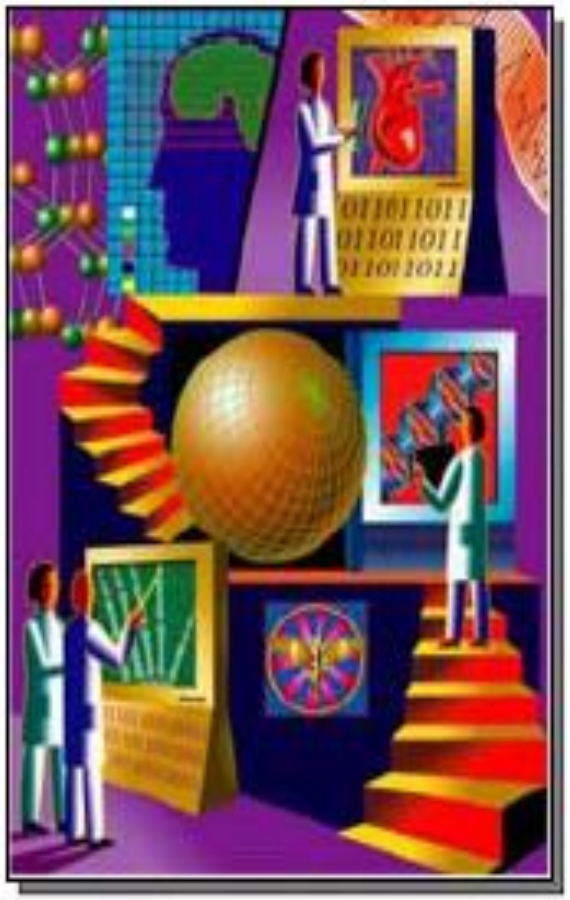
Mestrando: Marcus Vinícius Carneiro Teixeira
Orientador: Prof. Dr. Mauro Biajiz
Co-orientador: Prof. Dr. Ricardo Rodrigues Ciferri

Apresentação

- Introdução
- Anotação de Projetos Genoma
- Bancos de Dados de Genoma
- Ontologias e XML
- Ambientes de Anotação
- Ambiente BioFOX
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
  - Módulo Interface de Anotação
- Resultados e Conclusão

Introdução

- Motivação
  - Projetos Genoma geram um grande volume de dados.
  - Representação e armazenamento de dados biológicos.
  - Integração semântica de dados.
  - Ambiente de anotação Bio-TIM.



Introdução

- Objetivos
  - Ambiente de anotação BioFOX:
    - *Modelo de dados semi-estruturado;*
    - *Ontologias.*
  - Padronizar conceitos estabelecidos para o domínio;
  - Agregar semântica aos esquemas de dados e aos dados;
  - Criar bancos de dados flexíveis e apropriados para evolução de esquemas.

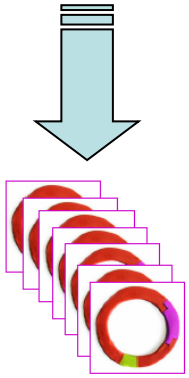
Introdução

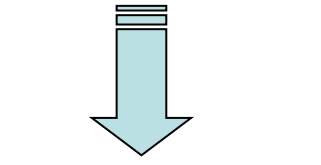
- Propostas:
  - Arquitetura para o Ambiente BioFOX;
  - Modelo Conceito-Compartilhado;
  - Interface para modelagem e geração de esquemas XML;
  - Interface para anotação manual.

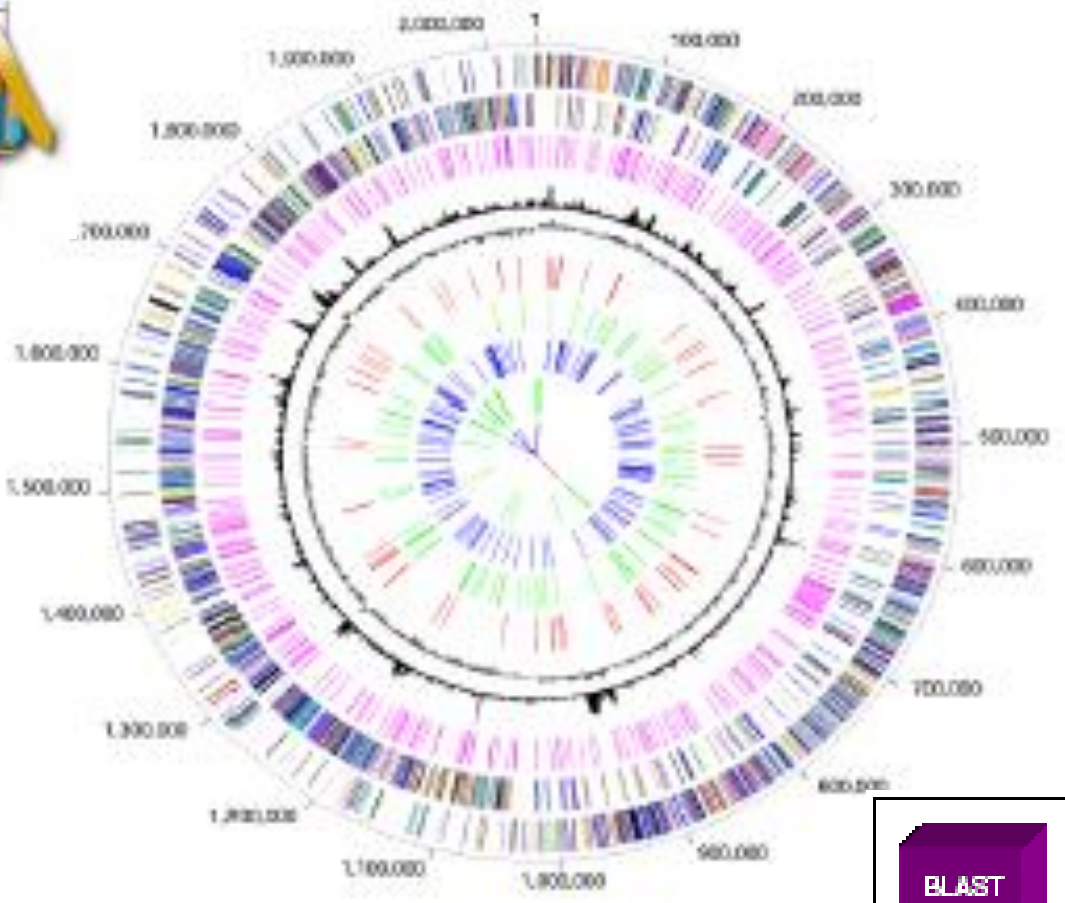
Apresentação

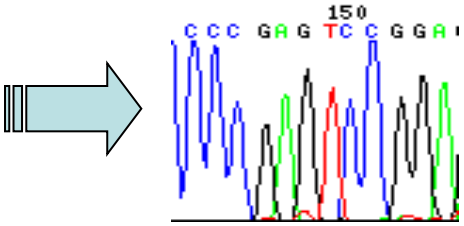
- Introdução
- Anotação de Projetos Genoma
- Bancos de Dados de Genoma
- Ontologias e XML
- Ambientes de Anotação
- Ambiente BioFOX
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
  - Módulo Interface de Anotação
- Resultados e Conclusão

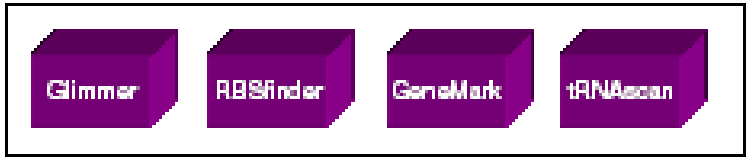
Anotação de Projetos Genoma











Anotação de Projetos Genoma

- Níveis de anotação genômica:

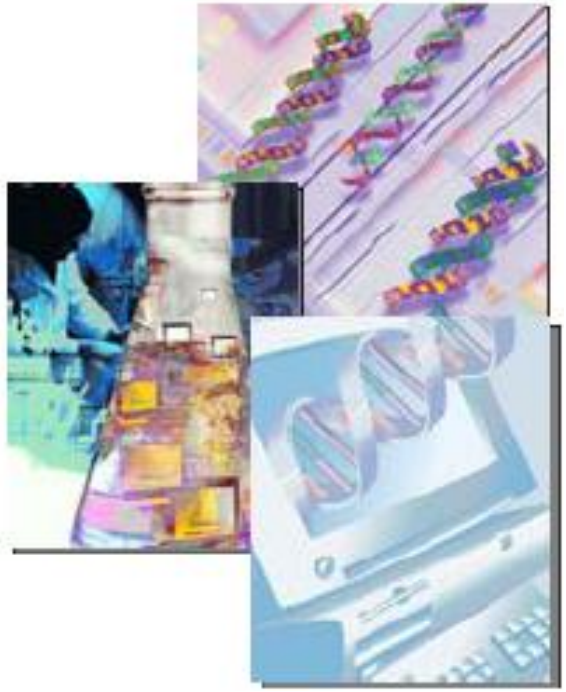
- Fontes de dados em ambientes de anotação:
  - Anotação importada;
  - Anotação automática;
  - Anotação manual.

Apresentação

- Introdução
- Anotação de Projetos Genoma
- Bancos de Dados de Genoma
- Ontologias e XML
- Ambientes de Anotação
- Ambiente BioFOX
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
  - Módulo Interface de Anotação
- Resultados e Conclusão

Bancos de Dados de Genoma

- Características:
  - Aplicações de bioinformática armazenam seus dados em formatos não padronizados;
  - Dados com estrutura irregular;
  - Necessidade de flexibilidade para evolução de esquema.



Bancos de Dados de Genoma

- É necessário tratar vários pontos importantes:

1. O controle semântico dos dados;
2. A definição do modelo de dados mais adequado;
3. As necessidades de processamento;
4. Os meios de acesso e o problema de integração de bancos de dados biológicos.

Apresentação

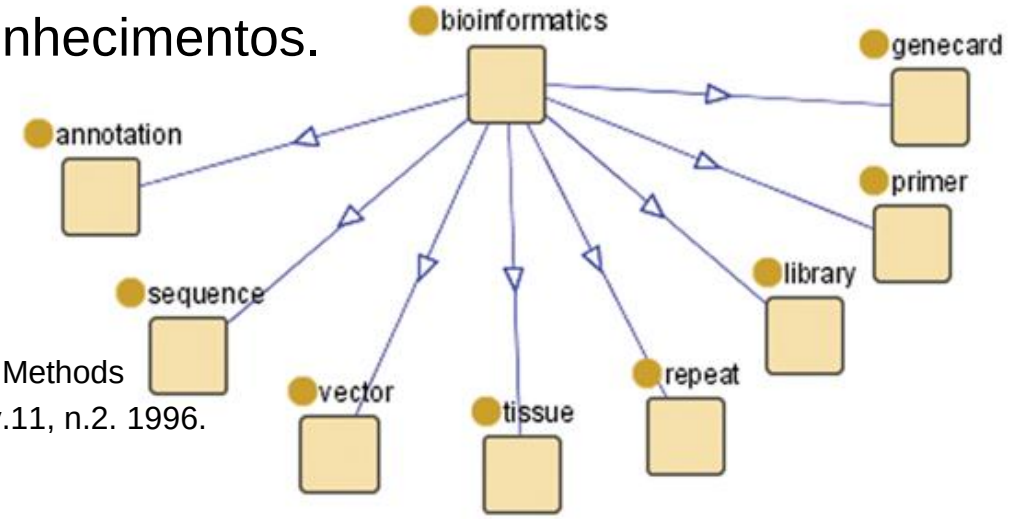
- Introdução
- Anotação de Projetos Genoma
- Bancos de Dados de Genoma
- Ontologias e XML
- Ambientes de Anotação
- Ambiente BioFOX
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
  - Módulo Interface de Anotação
- Resultados e Conclusão

Ontologias

1. Controle semântico dos dados.

- O que são ontologias?
  - Definição formal de conceitos pertinentes a um domínio, compartilhada por um grupo (Uschold e Gruninger, 1996).

- Para a bioinformática:
  - importante meio de padronização do vocabulário;
  - facilita a troca de informações;
  - agiliza a produção de conhecimentos.



Uschold, M. e M. Gruninger. Ontologies: Principles, Methods and Applications. Knowledge Engineering Review, v.11, n.2. 1996.

Dados Semi-Estruturados e XML

2. Modelo de dados adequado.

- Dados semi-estruturados apresentam:
  - representação estrutural heterogênea/irregular.
  - estrutura evolucionária.

- Bancos de Dados XML Nativos
  - usada na descrição de dados semi-estruturados.
  - a estrutura de um documento XML é determinante para o desempenho de acesso aos dados.

Apresentação

- Introdução
- Anotação de Projetos Genoma
- Bancos de Dados de Genoma
- Ontologias e XML
- Ambientes de Anotação
- Ambiente BioFOX
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
  - Módulo Interface de Anotação
- Resultados e Conclusão

Ambientes de Anotação

Ambientes de anotação	Armazenamento de dados	Tipos de anotação	Integração de dados	Vocabulário controlado	Ontologia
Apollo	SGBD Relacional	1, 2, 3	-	X	-
Ártemis	Formatos EMBL, GenBank e GFF	1, 2, 3	-	-	-
ASAP	SGBD Relacional	1, 2, 3	X	X	-
BASys	SGBD Relacional	1, 2	-	-	GO
BioNotes	SGBD Relacional Estendido	1, 2, 3	X	X	GO
ERGO	SGBD Relacional	1, 2, 3	X	-	ERGO
GARSA	SGBD Relacional	1, 2, 3	-	X	GO
GenDB	SGBD Relacional	1, 2, 3	-	-	GO
GeneQuiz	SGBD Relacional	1, 2	X	X	-
Genotator	Flat files (ACE)	2	-	-	-
Gopalacharyulu	SGBDs Relacional e Semi-estruturado	1, 2, 3	X	-	GO
MiGenes	SGBD Relacional	1, 2, 3	X	-	GO
PEDANT	SGBD Relacional	1, 2, 3	X	-	-
Bio-TIM	SGBD Relacional	1, 2, 3	X	-	-

Ambientes de Anotação

Ambientes de anotação	Armazenamento de dados	Tipos de anotação	Integração de dados	Vocabulário controlado	Ontologia
Apollo	SGBD Relacional	1, 2, 3	-	X	-
Ártemis	Formatos EMBL, GenBank e GFF	1, 2, 3	-	-	-
ASAP	SGBD Relacional	1, 2, 3	X	X	-
BASys	SGBD Relacional	1, 2	-	-	GO
BioNotes	SGBD Relacional Estendido	1, 2, 3	X	X	GO
ERGO	SGBD Relacional	1, 2, 3	X	-	ERGO
GARSA	SGBD Relacional	1, 2, 3	-	X	GO
GenDB	SGBD Relacional	1, 2, 3	-	-	GO
GeneQuiz	SGBD Relacional	1, 2	X	X	-
Genotator	Flat files (ACE)	2	-	-	-
Gopalacharyulu	SGBDs Relacional e Semi-estruturado	1, 2, 3	X	-	GO
MiGenes	SGBD Relacional	1, 2, 3	X	-	GO
PEDANT	SGBD Relacional	1, 2, 3	X	-	-
Bio-TIM	SGBD Relacional	1, 2, 3	X	-	-

Ambientes de Anotação

Ambientes de anotação	Armazenamento de dados	Tipos de anotação	Integração de dados	Vocabulário controlado	Ontologia
Apollo	SGBD Relacional	1, 2, 3	-	X	-
Ártemis	Formatos EMBL, GenBank e GFF	1, 2, 3	-	-	-
ASAP	SGBD Relacional	1, 2, 3	X	X	-
BASys	SGBD Relacional	1, 2	-	-	GO
BioNotes	SGBD Relacional Estendido	1, 2, 3	X	X	GO
ERGO	SGBD Relacional	1, 2, 3	X	-	ERGO
GARSA	SGBD Relacional	1, 2, 3	-	X	GO
GenDB	SGBD Relacional	1, 2, 3	-	-	GO
GeneQuiz	SGBD Relacional	1, 2	X	X	-
Genotator	Flat files (ACE)	2	-	-	-
Gopalacharyulu	SGBDs Relacional e Semi-estruturado	1, 2, 3	X	-	GO
MiGenes	SGBD Relacional	1, 2, 3	X	-	GO
PEDANT	SGBD Relacional	1, 2, 3	X	-	-
Bio-TIM	SGBD Relacional	1, 2, 3	X	-	-

Ambientes de Anotação

Ambientes de anotação	Armazenamento de dados	Tipos de anotação	Integração de dados	Vocabulário controlado	Ontologia
Apollo	SGBD Relacional	1, 2, 3	-	X	-
Ártemis	Formatos EMBL, GenBank e GFF	1, 2, 3	-	-	-
ASAP	SGBD Relacional	1, 2, 3	X	X	-
BASys	SGBD Relacional	1, 2	-	-	GO
BioNotes	SGBD Relacional Estendido	1, 2, 3	X	X	GO
ERGO	SGBD Relacional	1, 2, 3	X	-	ERGO
GARSA	SGBD Relacional	1, 2, 3	-	X	GO
GenDB	SGBD Relacional	1, 2, 3	-	-	GO
GeneQuiz	SGBD Relacional	1, 2	X	X	-
Genotator	Flat files (ACE)	2	-	-	-
Gopalacharyulu	SGBDs Relacional e Semi-estruturado	1, 2, 3	X	-	GO
MiGenes	SGBD Relacional	1, 2, 3	X	-	GO
PEDANT	SGBD Relacional	1, 2, 3	X	-	-
Bio-TIM	SGBD Relacional	1, 2, 3	X	-	-

Ambientes de Anotação

Ambientes de anotação	Armazenamento de dados	Tipos de anotação	Integração de dados	Vocabulário controlado	Ontologia
Apollo	SGBD Relacional	1, 2, 3	-	X	-
Ártemis	Formatos EMBL, GenBank e GFF	1, 2, 3	-	-	-
ASAP	SGBD Relacional	1, 2, 3	X	X	-
BASys	SGBD Relacional	1, 2	-	-	GO
BioNotes	SGBD Relacional Estendido	1, 2, 3	X	X	GO
ERGO	SGBD Relacional	1, 2, 3	X	-	ERGO
GARSA	SGBD Relacional	1, 2, 3	-	X	GO
GenDB	SGBD Relacional	1, 2, 3	-	-	GO
GeneQuiz	SGBD Relacional	1, 2	X	X	-
Genotator	Flat files (ACE)	2	-	-	-
Gopalacharyulu	SGBDs Relacional e Semi-estruturado	1, 2, 3	X	-	GO
MiGenes	SGBD Relacional	1, 2, 3	X	-	GO
PEDANT	SGBD Relacional	1, 2, 3	X	-	-
Bio-TIM	SGBD Relacional	1, 2, 3	X	-	-

Ambientes de Anotação

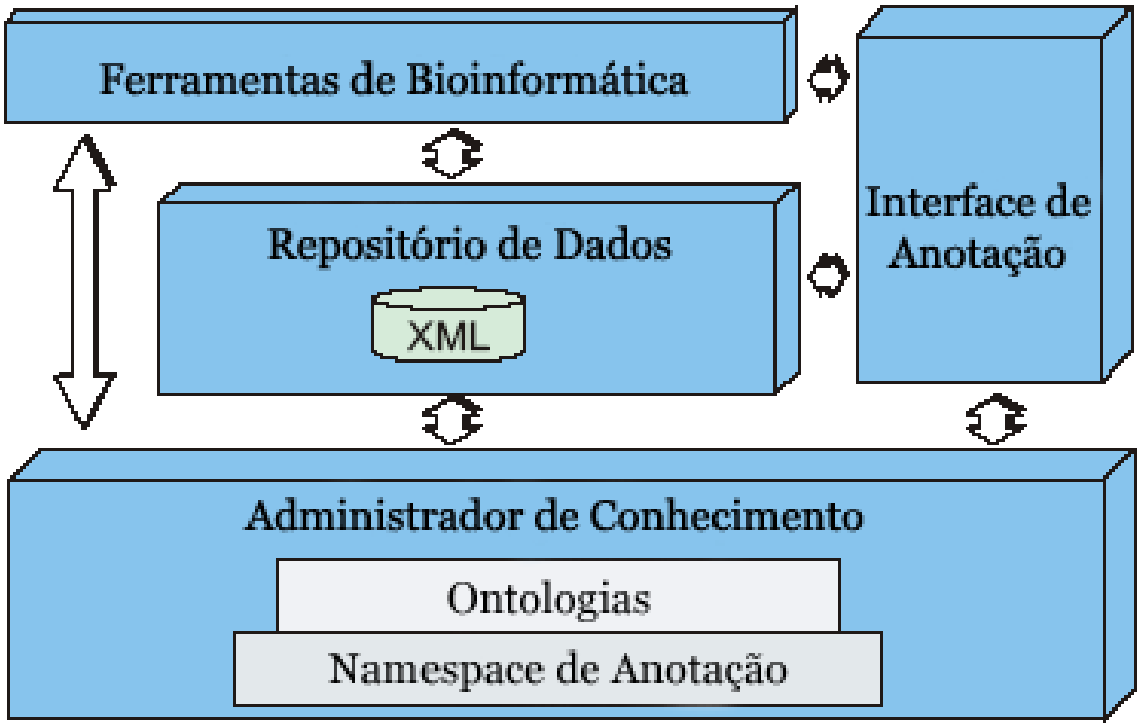
Ambientes de anotação	Armazenamento de dados	Tipos de anotação	Integração de dados	Vocabulário controlado	Ontologia
Apollo	SGBD Relacional	1, 2, 3	-	X	-
Ártemis	Formatos EMBL, GenBank e GFF	1, 2, 3	-	-	-
ASAP	SGBD Relacional	1, 2, 3	X	X	-
BASys	SGBD Relacional	1, 2	-	-	GO
BioNotes	SGBD Relacional Estendido	1, 2, 3	X	X	GO
ERGO	SGBD Relacional	1, 2, 3	X	-	ERGO
GARSA	SGBD Relacional	1, 2, 3	-	X	GO
GenDB	SGBD Relacional	1, 2, 3	-	-	GO
GeneQuiz	SGBD Relacional	1, 2	X	X	-
Genotator	Flat files (ACE)	2	-	-	-
Gopalacharyulu	SGBDs Relacional e Semi-estruturado	1, 2, 3	X	-	GO
MiGenes	SGBD Relacional	1, 2, 3	X	-	GO
PEDANT	SGBD Relacional	1, 2, 3	X	-	-
BioFOX	SGBD Semi-estruturado e DW Relacional	1, 2, 3	X	X	GO e SO

Apresentação

- Introdução
- Anotação de Projetos Genoma
- Bancos de Dados de Genoma
- Ontologias e XML
- Ambientes de Anotação
- Ambiente BioFOX
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
  - Módulo Interface de Anotação
- Resultados e Conclusão

Ambiente BioFOX

- Arquitetura:
  - Módulo Ferramentas de Bioinformática (MFB)
  - Módulo Repositório de Dados (MRD)
  - Módulo Interface de Anotação (MIA)
  - Módulo Administrador de Conhecimento (MAC)



Apresentação

- Introdução
- Anotação de Projetos Genoma
- Bancos de Dados de Genoma
- Ontologias e XML
- Ambientes de Anotação
- Ambiente BioFOX
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
  - Módulo Interface de Anotação
- Resultados e Conclusão

Administrador de Conhecimento (MAC)

- Onde as ontologias e o *Namespace* de Anotação são definidos.
- Ontologia de aplicação e *Namespace* representam o mesmo conjunto de dados.
- Deve conhecer e organizar todas os dados de anotação provenientes dos demais módulos

Apresentação

- Introdução
- Anotação de Projetos Genoma
- Bancos de Dados de Genoma
- Ontologias e XML
- Ambientes de Anotação
- Ambiente BioFOX
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
  - Módulo Interface de Anotação
- Resultados e Conclusão

Repositório de Dados (MRD)

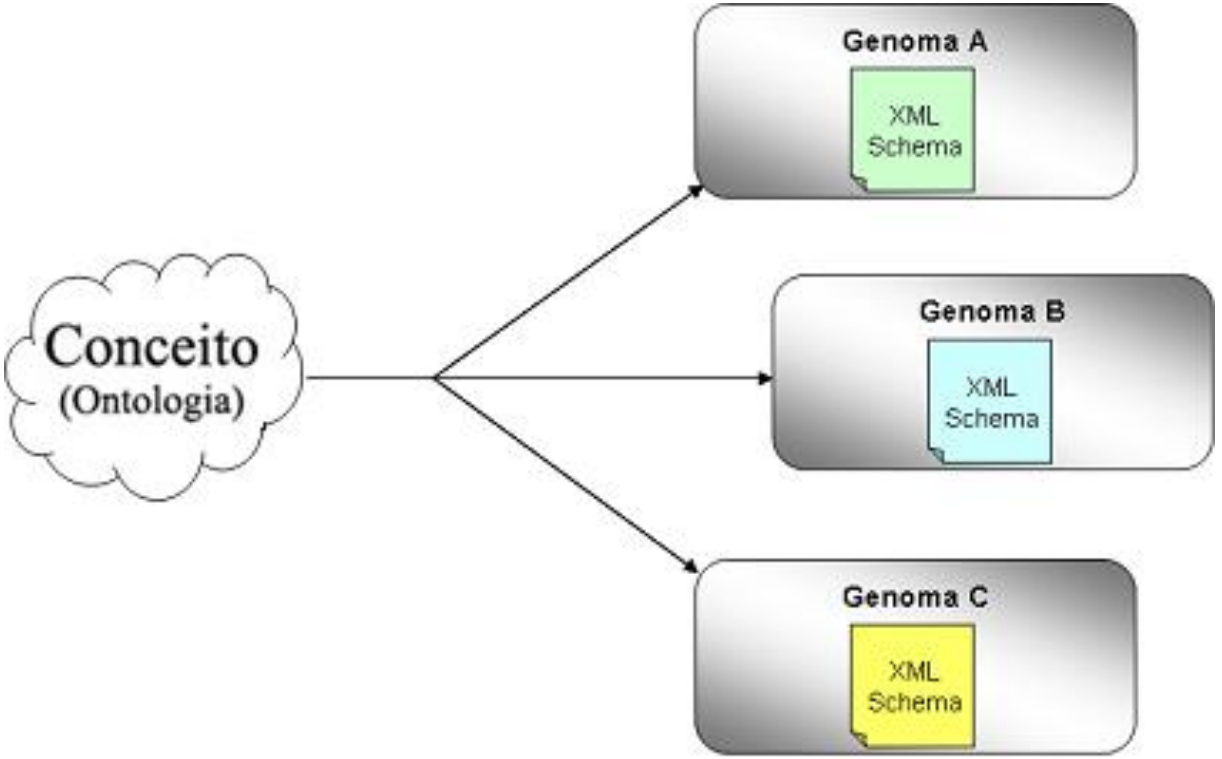
- Estrutura para o armazenamento de anotações.

- SGBD XML nativo (Tamino XML Server)
  - Esquemas flexíveis.
  - Bancos de dados inter-operáveis.
  - Desenvolvimento de uma interface para o projeto de banco de dados XML.
  - Semântica associada aos esquemas de dados (*conceito-compartilhado*).

tamino

tamino XML Server

Conceito-Compartilhado



3. Necessidades de processamento.

Esquemas diferentes!

4. Integração de dados.

Mesma semântica!

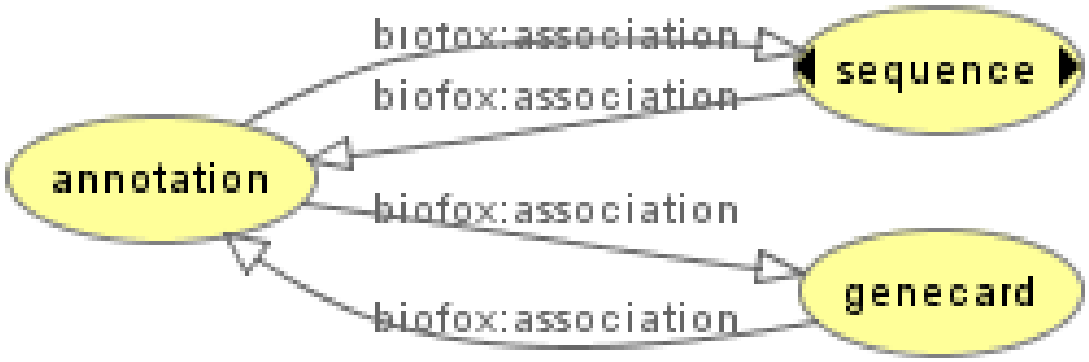
Projetos genoma de um mesmo domínio de pesquisa não necessariamente trabalham com o mesmo conjunto de anotações.

Ontologias

- Ontologia de aplicação:
  - Associação;
  - Agrupamento;
  - Parte de.

Ontologias

- Ontologia de aplicação:
  - Associação;
  - Agrupamento;
  - Parte de.

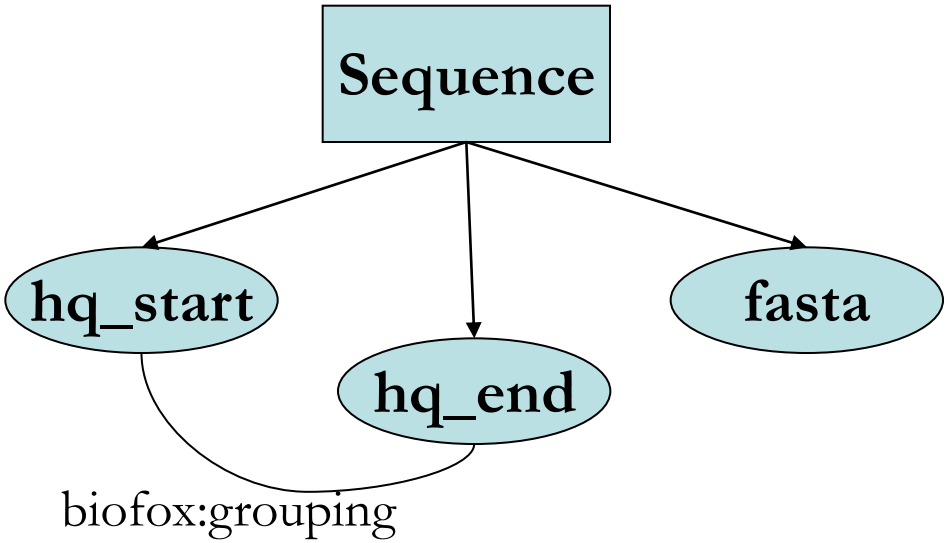



```
<sequence>
   <annotation/>
</sequence>
```

```
<annotation>
   <sequence/>
</annotation>
```

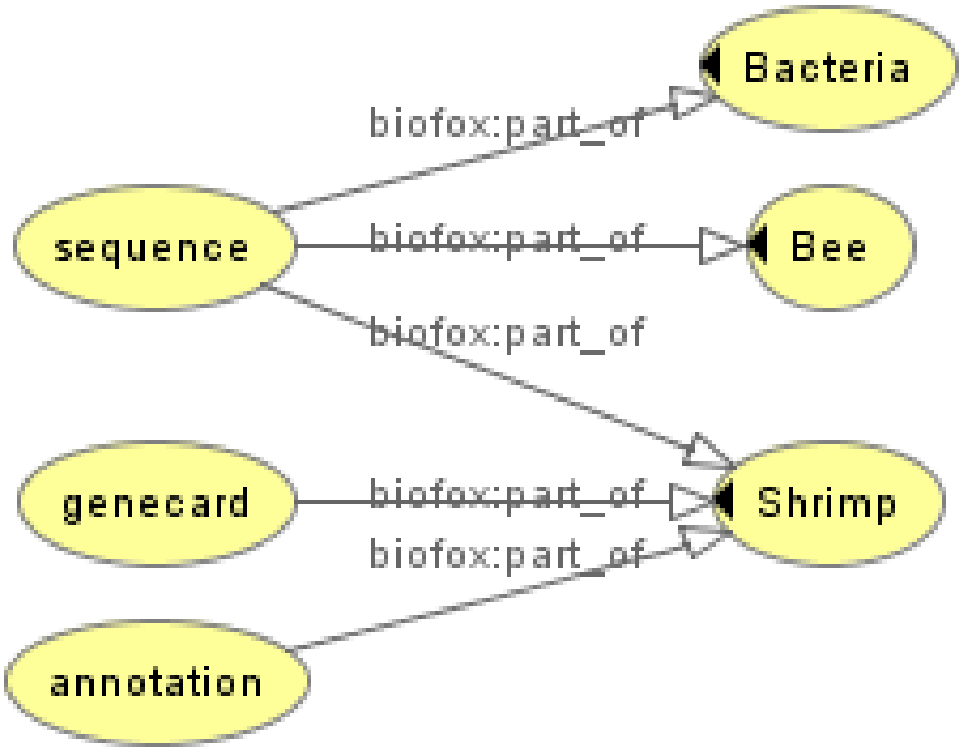
Ontologias

- Ontologia de aplicação:
  - Associação;
  - Agrupamento;
  - Parte de.

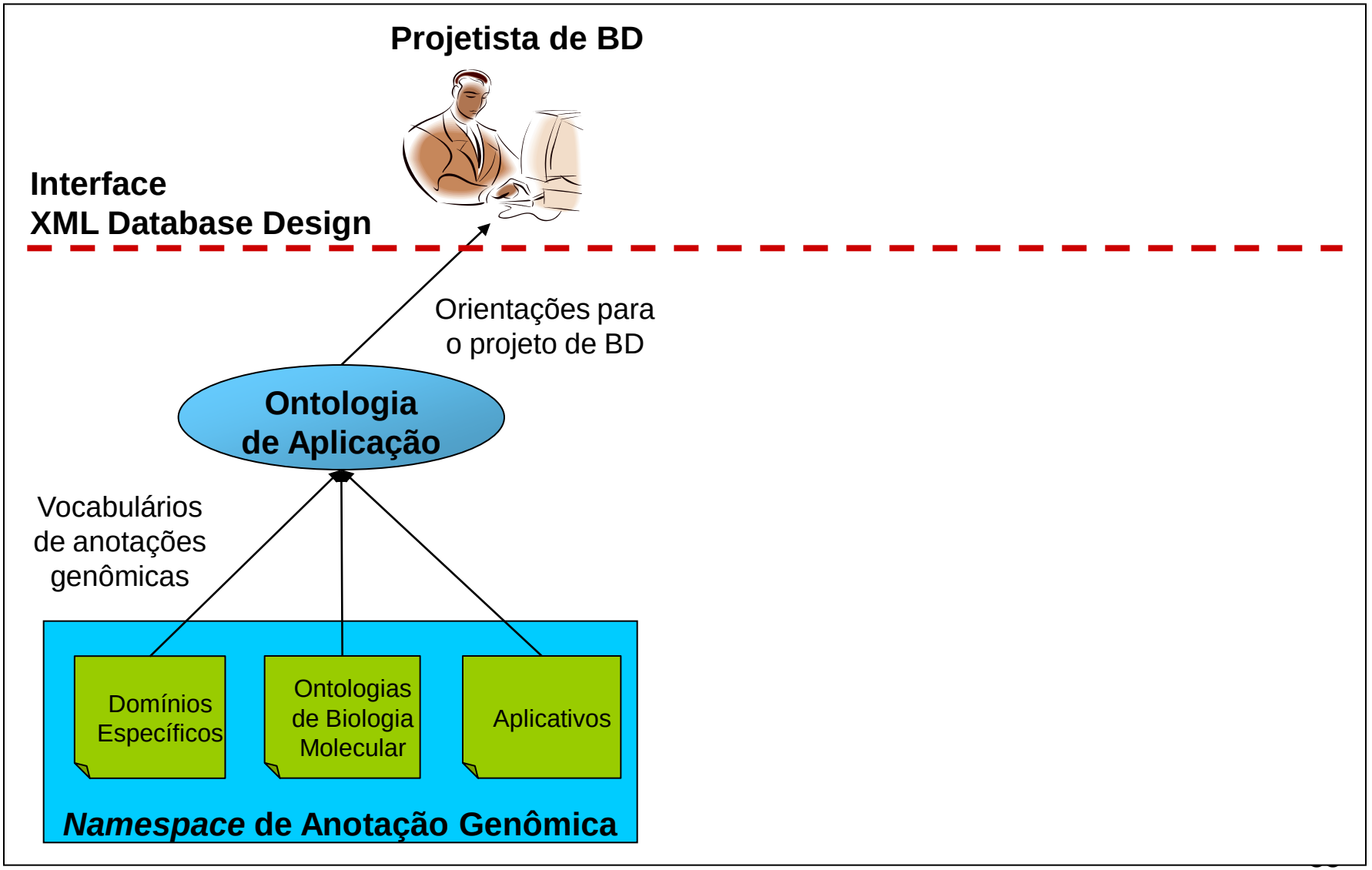


Ontologias

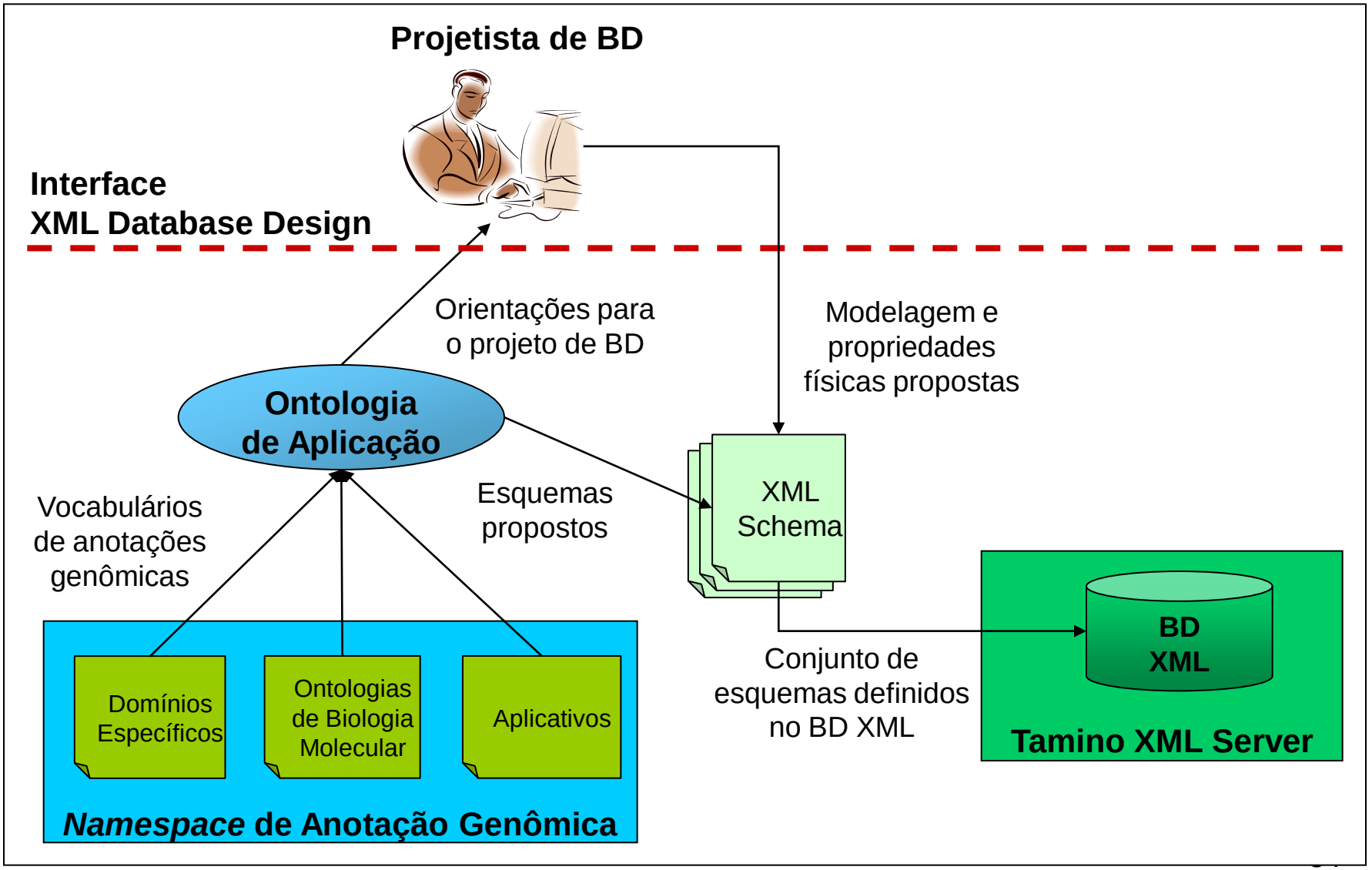
- Ontologia de aplicação:
  - Associação;
  - Agrupamento;
  - Parte de.



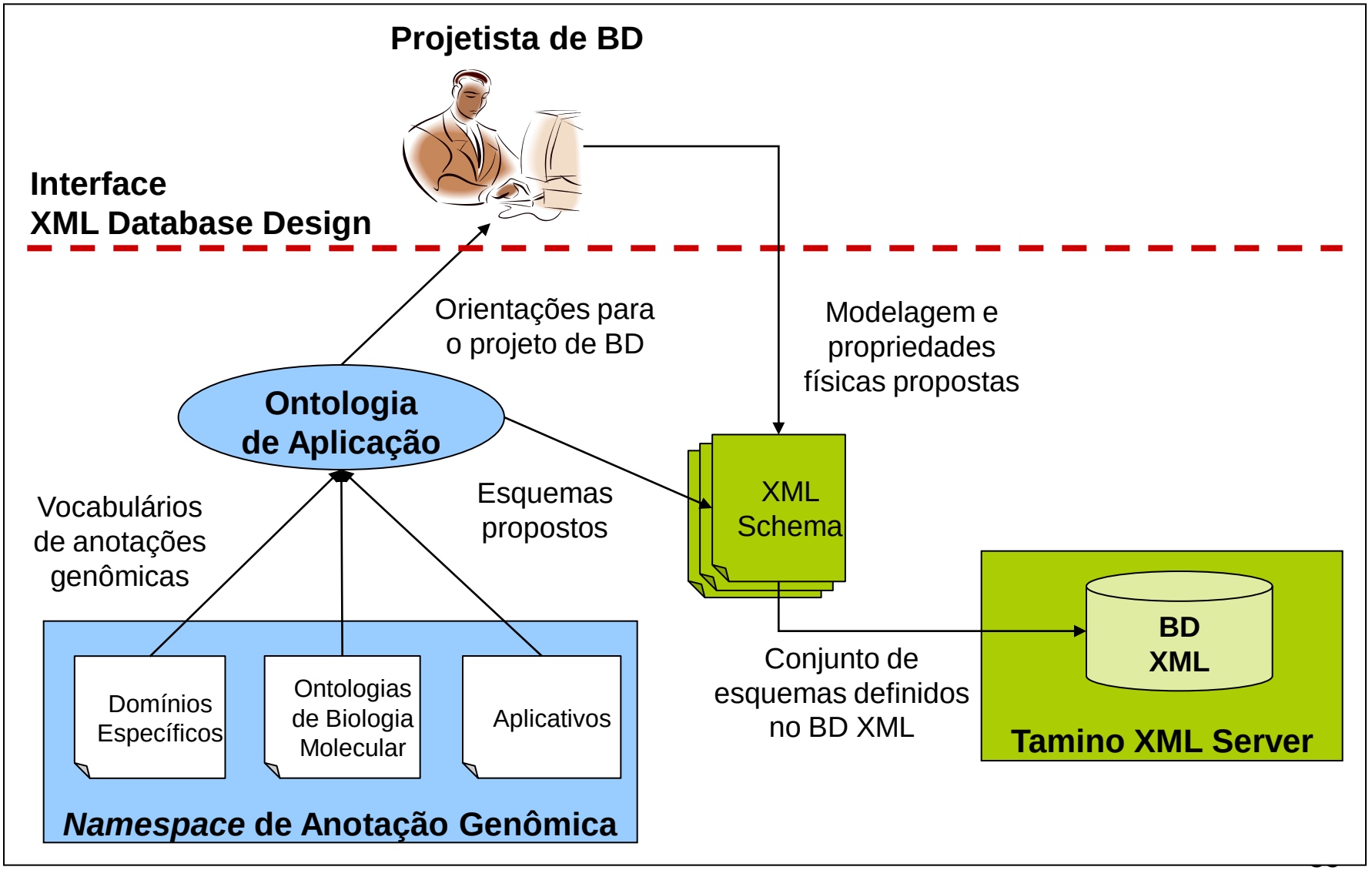
Repositório de Dados (MRD)



Repositório de Dados (MRD)



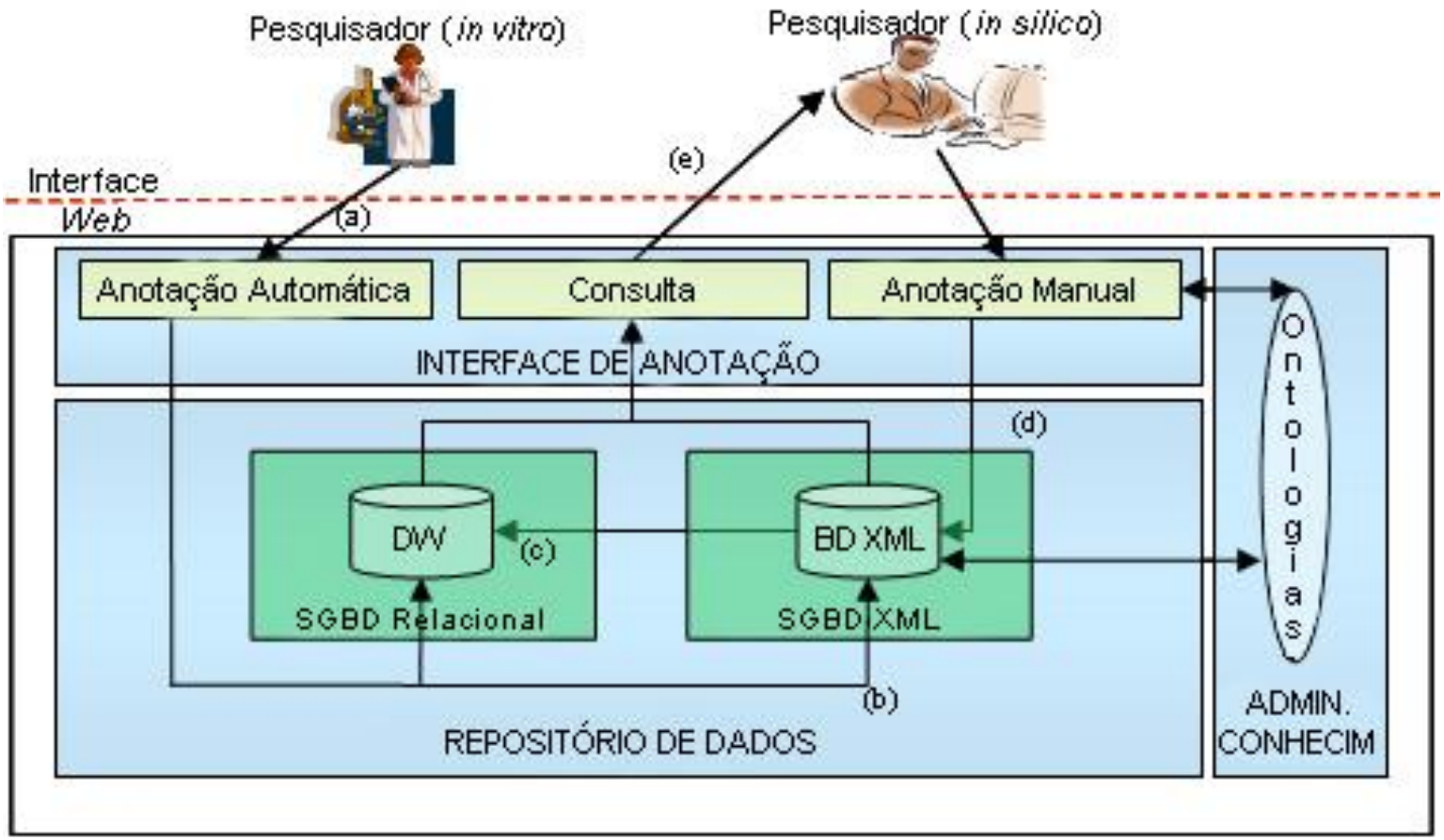
Repositório de Dados (MRD)



Apresentação

- Introdução
- Anotação de Projetos Genoma
- Ontologias
- Bancos de Dados de Genoma e XML
- Ambientes de Anotação
- Ambiente BioFOX
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
  - Módulo Interface de Anotação
- Resultados e Conclusão

Interface com Pesquisadores



Interfaces de Anotação Manual

- Como uma ontologia pode ajudar o pesquisador em sua anotação manual?
- Quais são as necessidades dos pesquisadores?
- Quais dificuldades com interfaces de programas convencionais?
- Que funcionalidades podem ser adicionadas?

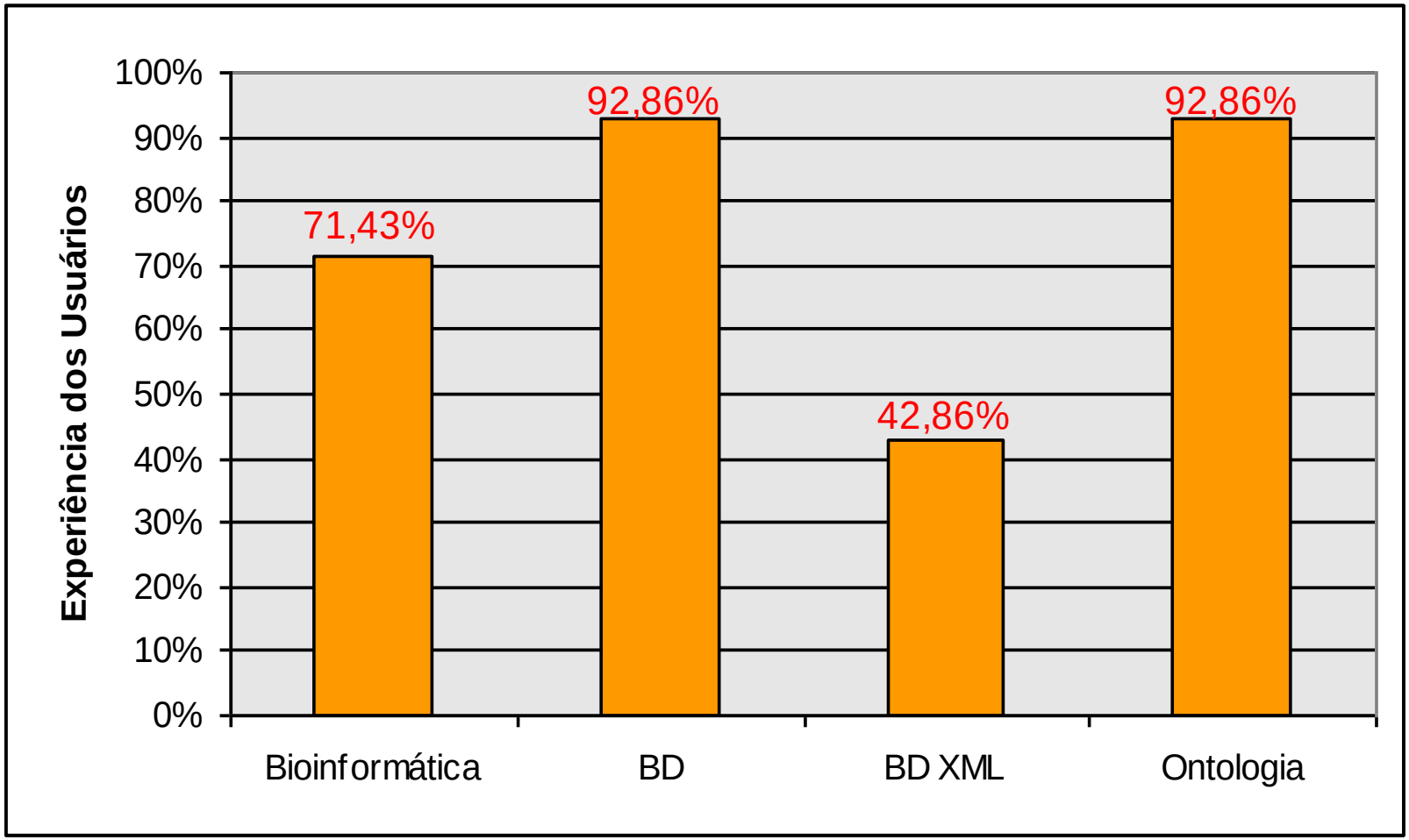
Interfaces de Anotação Manual

- Funcionalidades implementadas:
  - Auto-completar;
  - Sinônimos;
  - Exemplos;
  - *Definição de novos campos (em construção)*.

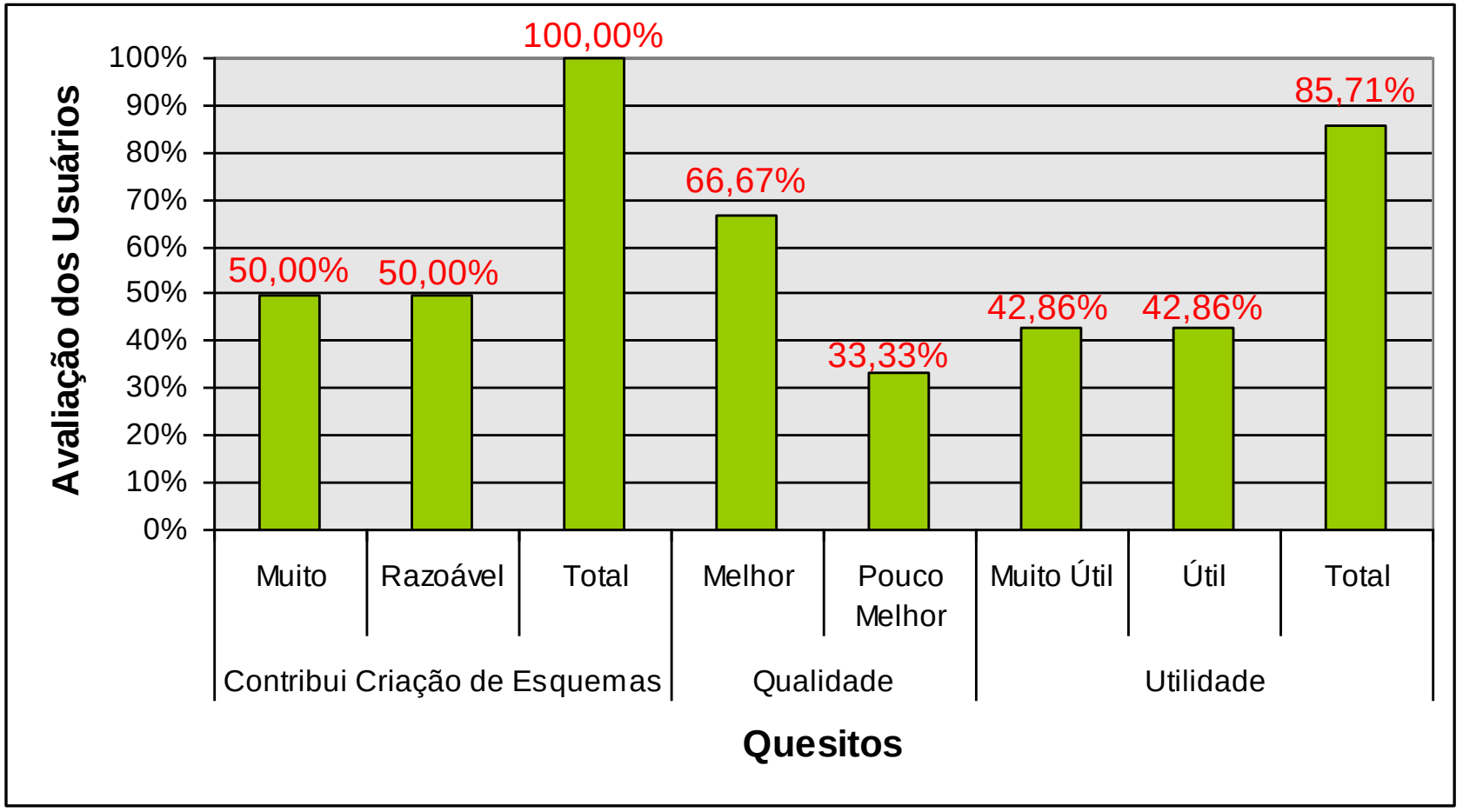
Apresentação

- Introdução
- Anotação de Projetos Genoma
- Ontologias
- Bancos de Dados de Genoma e XML
- Ambientes de Anotação
- Ambiente BioFOX
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
  - Módulo Interface de Anotação
- Resultados e Conclusão

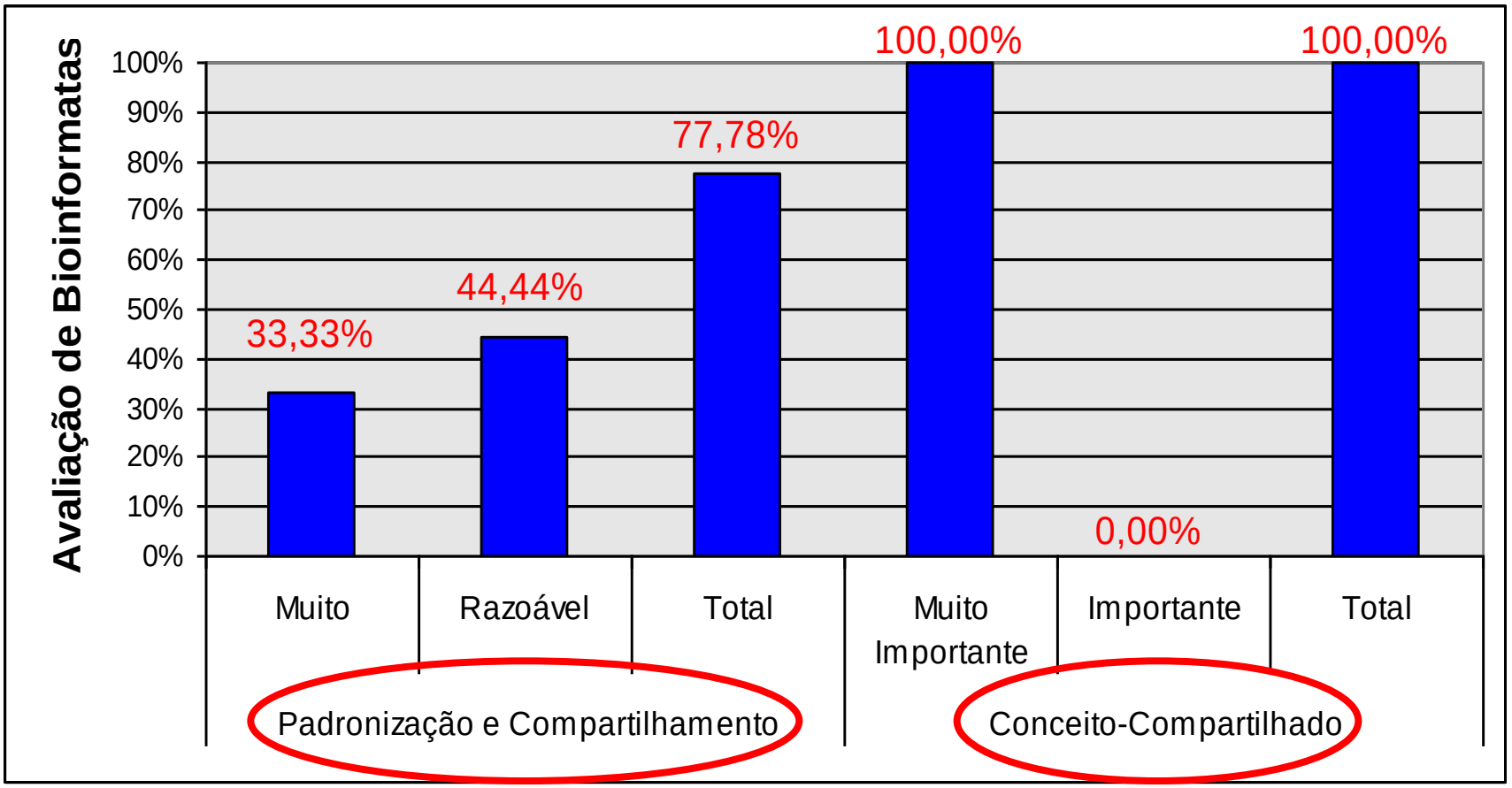
Experiência dos Usuários



Avaliação dos Usuários



Avaliação dos Usuários



Avaliação dos Usuários

- Aspectos positivos citados pelos usuários:
  - Visualização gráfica dos elementos em uma interface intuitiva e de fácil usabilidade;
  - Recomendação de termos e propriedades;
  - Vocabulário comum entre o projetista do banco de dados e biólogos especialistas no domínio;
  - Estabelecimento de regras do domínio, o que permite alertar para eventuais erros conceituais do projetista;

Avaliação dos Usuários

- Aspectos positivos citados pelos usuários:
  - Diminuição da heterogeneidade semântica entre aplicações de mesmo domínio;
  - Diminuição no tempo gasto para desenvolvimento do projeto de banco de dados e também para criação de esquemas XML;
  - Diminuição da disparidade semântica do esquema, reduzindo-se o custo de futuros esforços de integração.

Contribuições

- Proposta de uma arquitetura para um ambiente de anotação;
- Desenvolvimento e implementação de uma interface para a modelagem conceitual de domínios complexos, como o de biologia molecular, e criação de esquemas de dados XML;
- Criação de esquemas de dados estruturalmente independentes, mas com semântica associada por meio de uma ontologia;
- Proposta do modelo de integração conceito-compartilhado, o qual explora as características de esquemas de dados XML associados a uma ontologia;

Contribuições

- Desenvolvimento de uma ontologia de aplicação contendo definições de conceitos e regras de domínio, com o objetivo de guiar e auxiliar a modelagem conceitual de dados;

- Desenvolvimento de um namespace de vocabulários XML para a anotação de projetos genoma.

Trabalhos Futuros

- Exploração da integração de dados a partir do modelo de conceito-compartilhado apresentado neste trabalho;
- Componentização da interface de anotação manual;
- Testes de desempenho de bancos de dados XML quando aplicados em um projeto genoma;
- Expansão dos vocabulários de domínios já definidos e a inclusão de novos domínios;

Trabalhos Futuros

- Aplicação da interface XML Database Design a diferentes domínios de conhecimento, além do domínio de biologia molecular;

- Melhoria da interface XML Database Design.

26 de Maio de 2008.

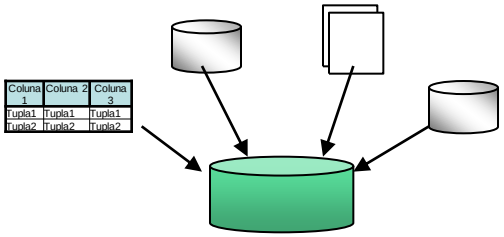
Anexos

Defesa de Mestrado

26 de Maio de 2008.

Anotação de Projetos Genoma

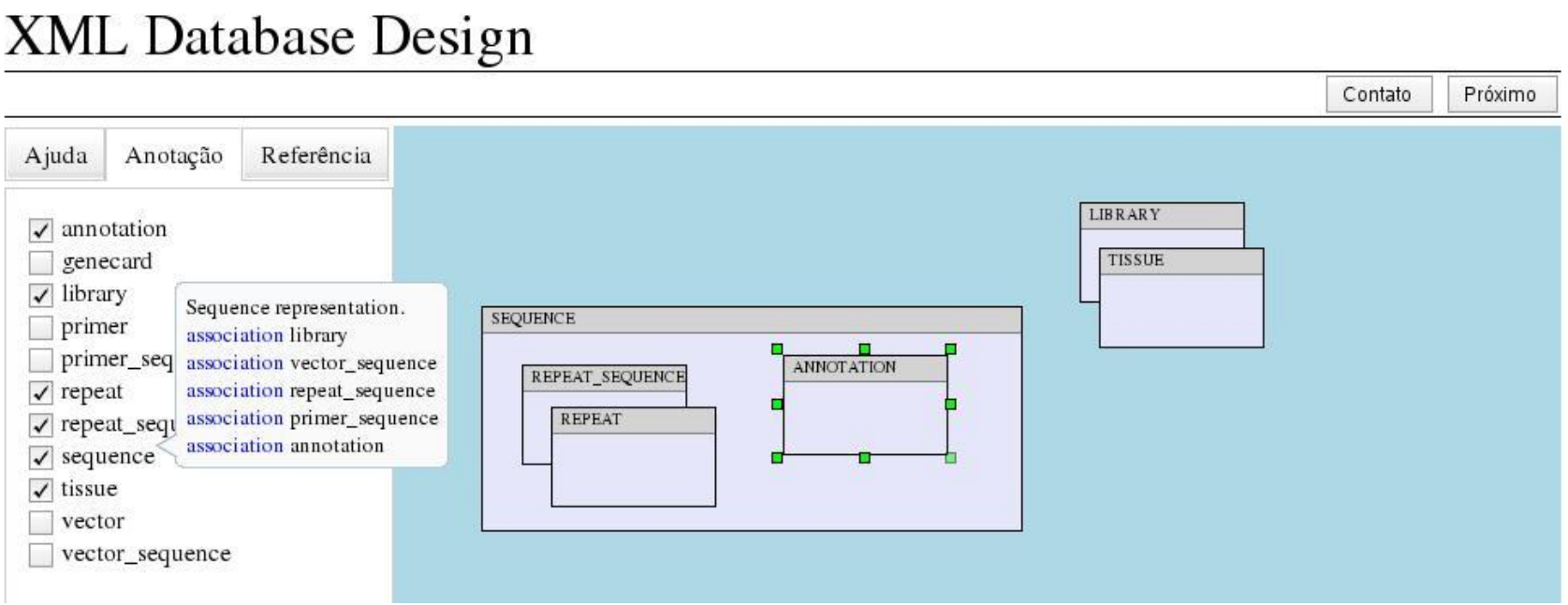
- Anotação importada

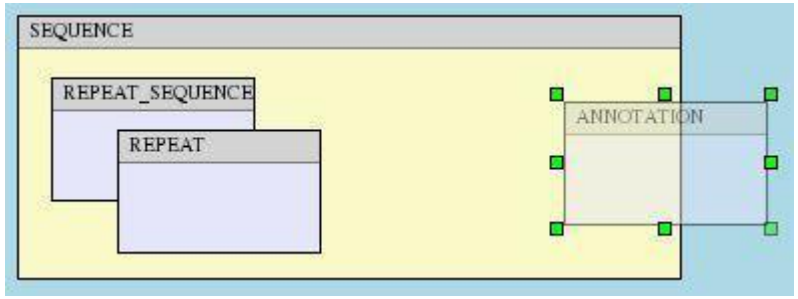


- Anotação automática

- Anotação manual

XML Database Design

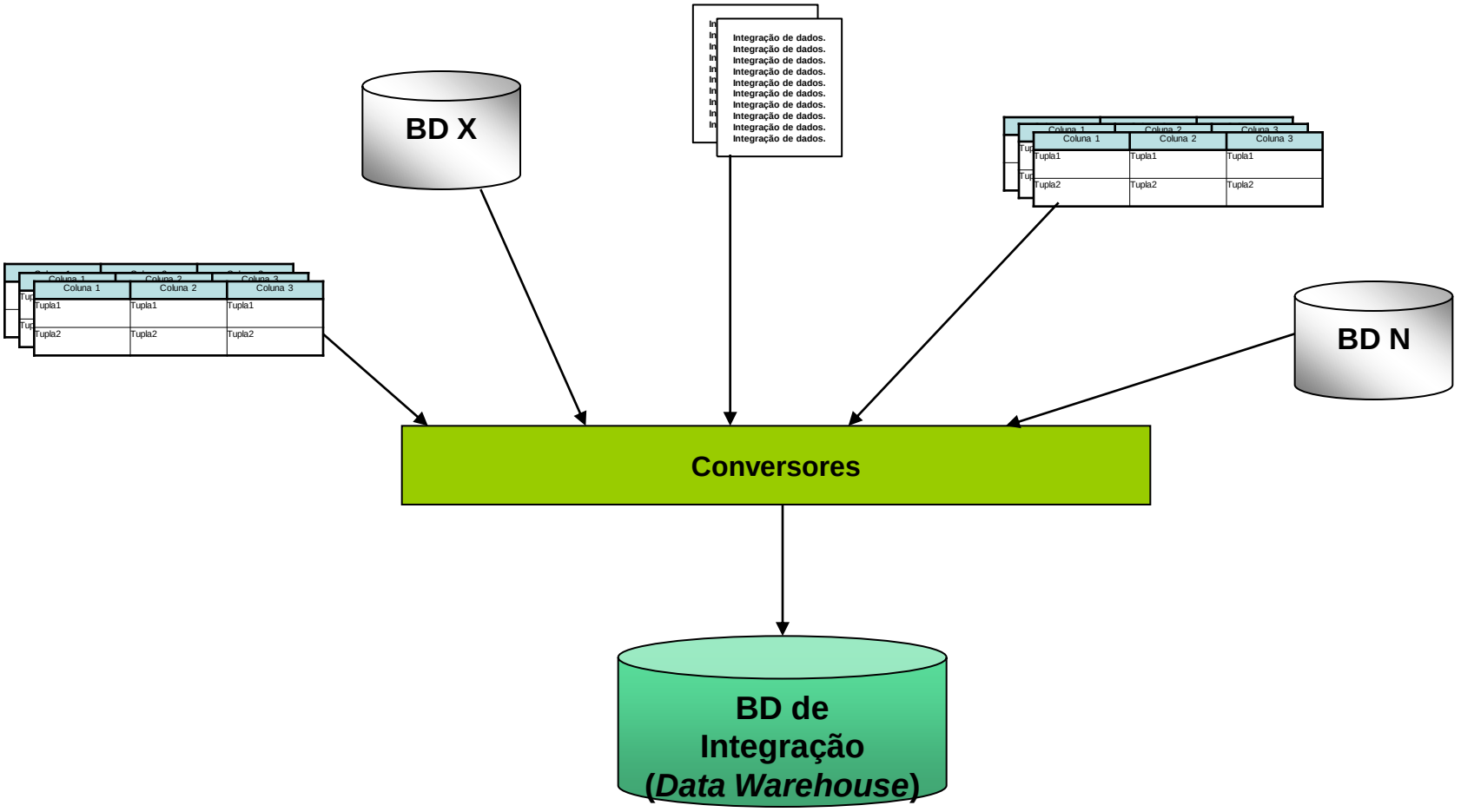




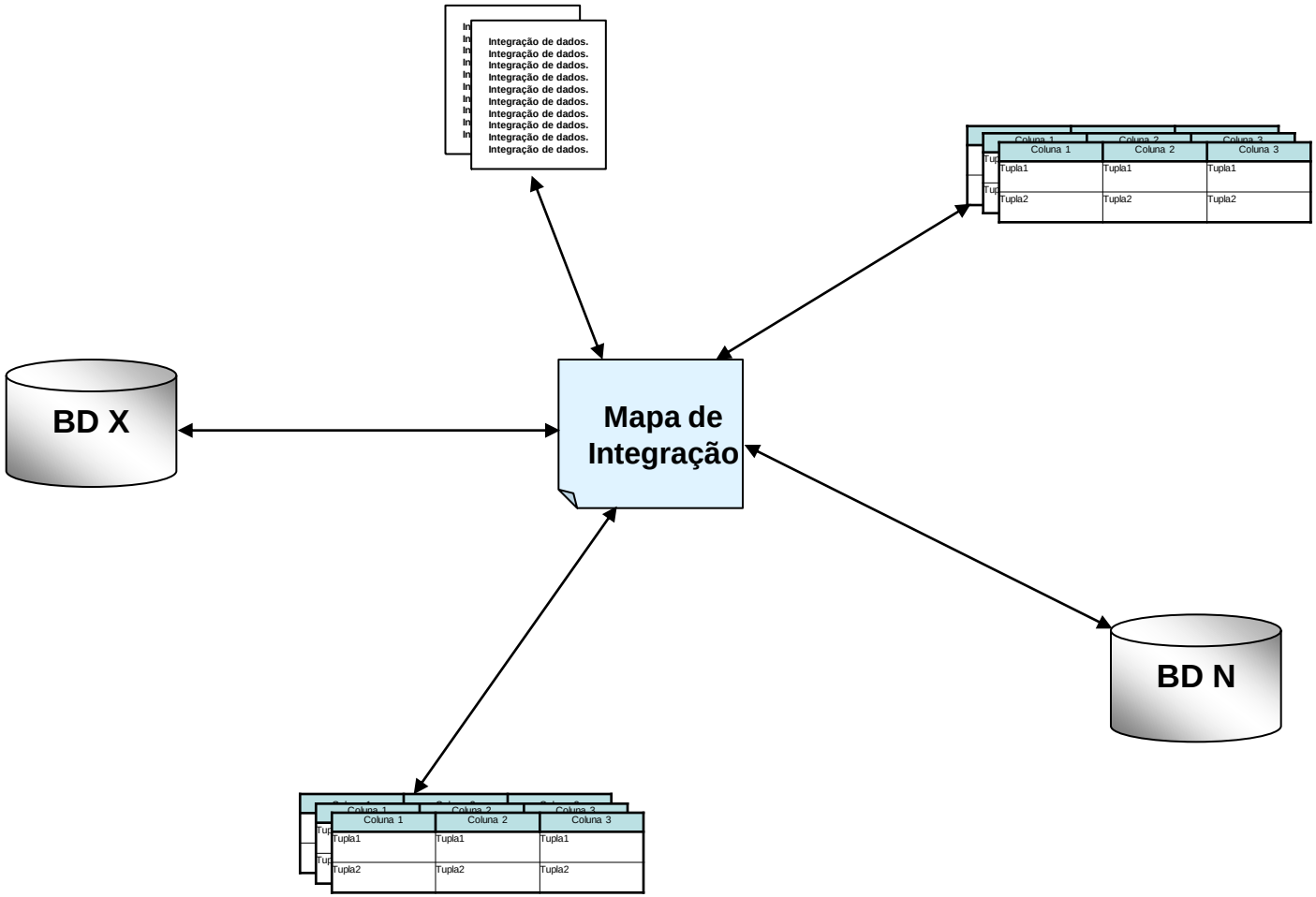
Apresentação

- Introdução
- Anotação de Projetos Genoma
- Ontologias
- Bancos de Dados de Genoma e XML
- Arquitetura Proposta
  - Módulo Administrador de Conhecimento
  - Módulo Repositório de Dados
- Interfaces para Anotação Manual
- Integração de Dados

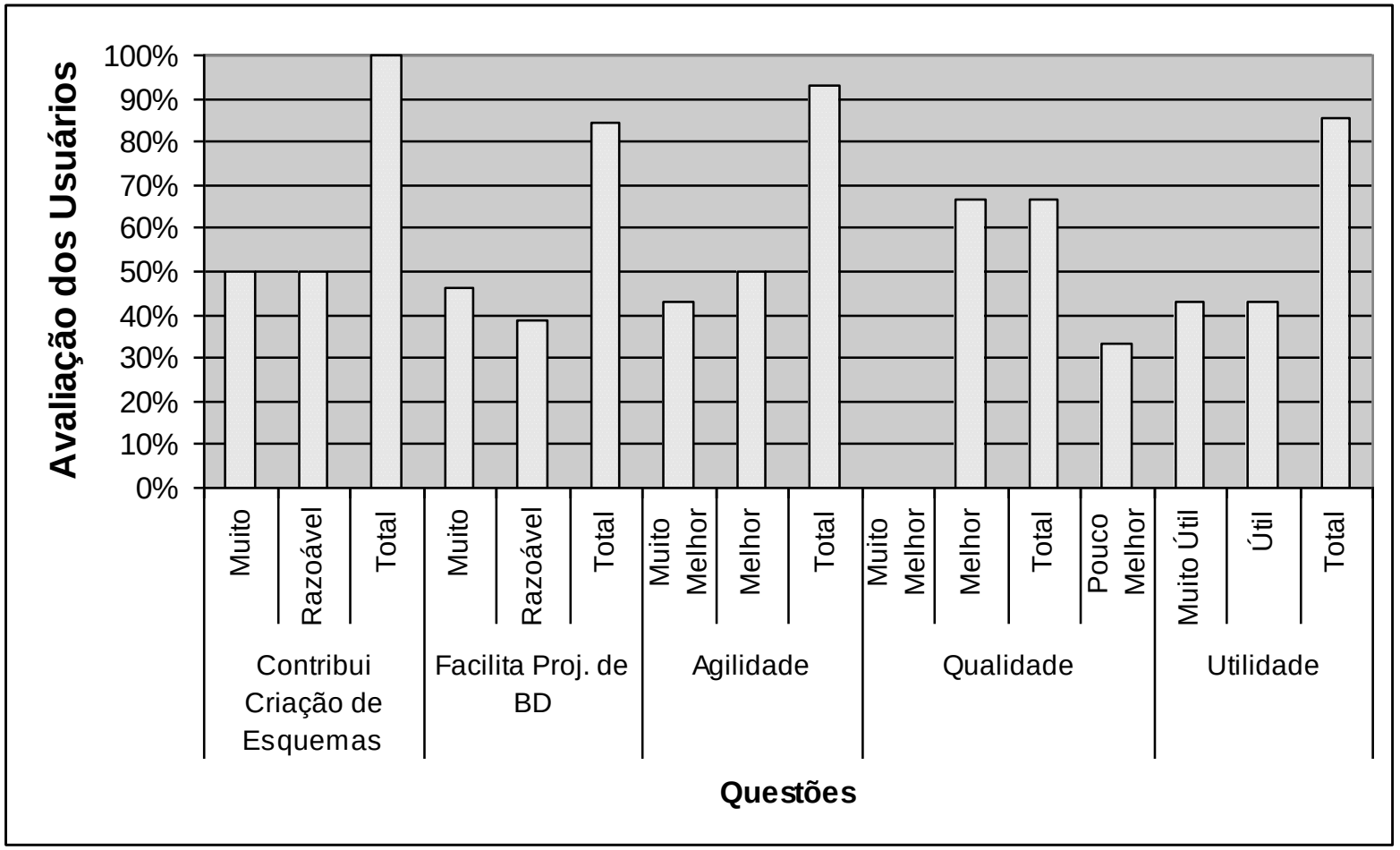
Integração Física



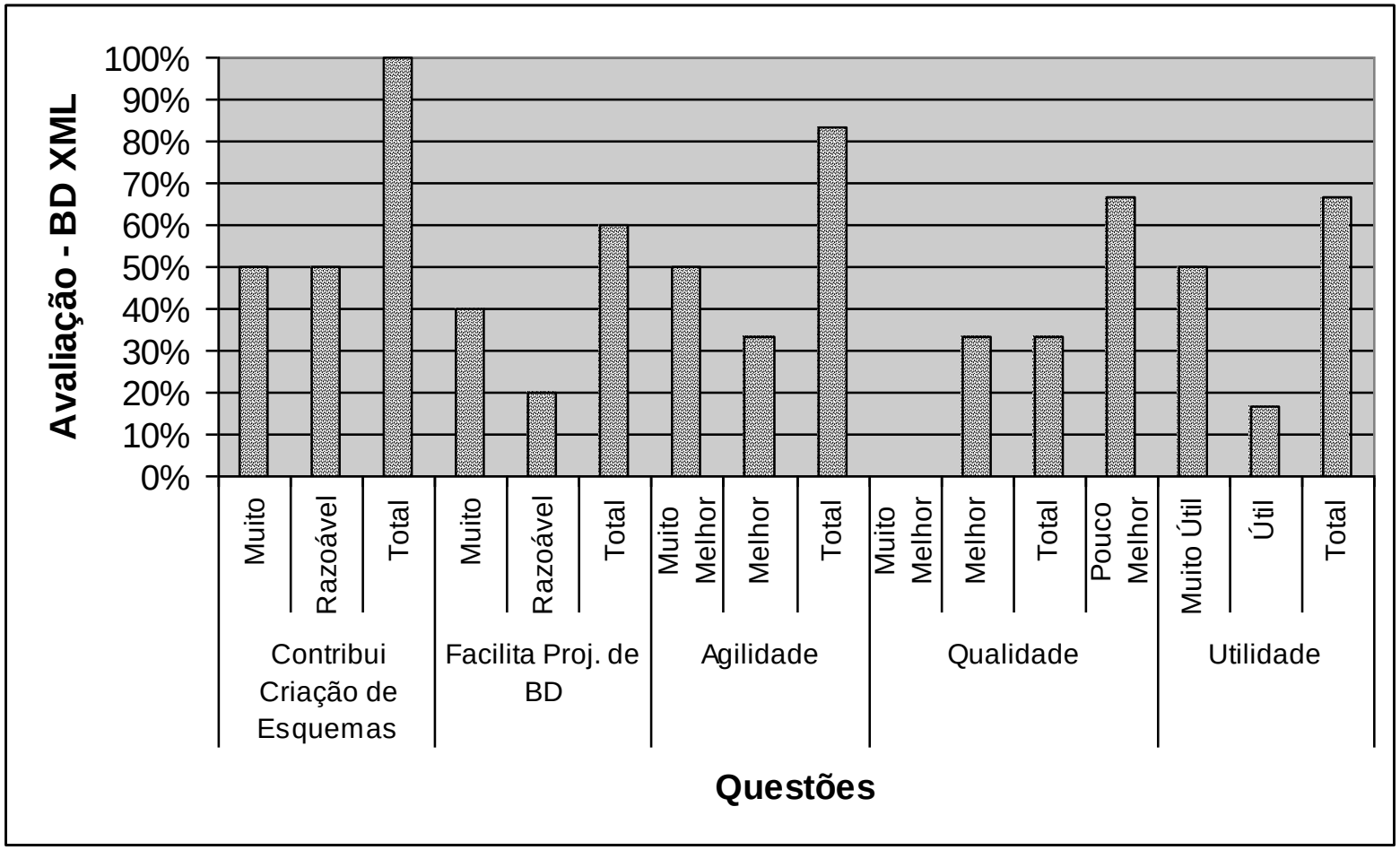
Integração por Mapeamento



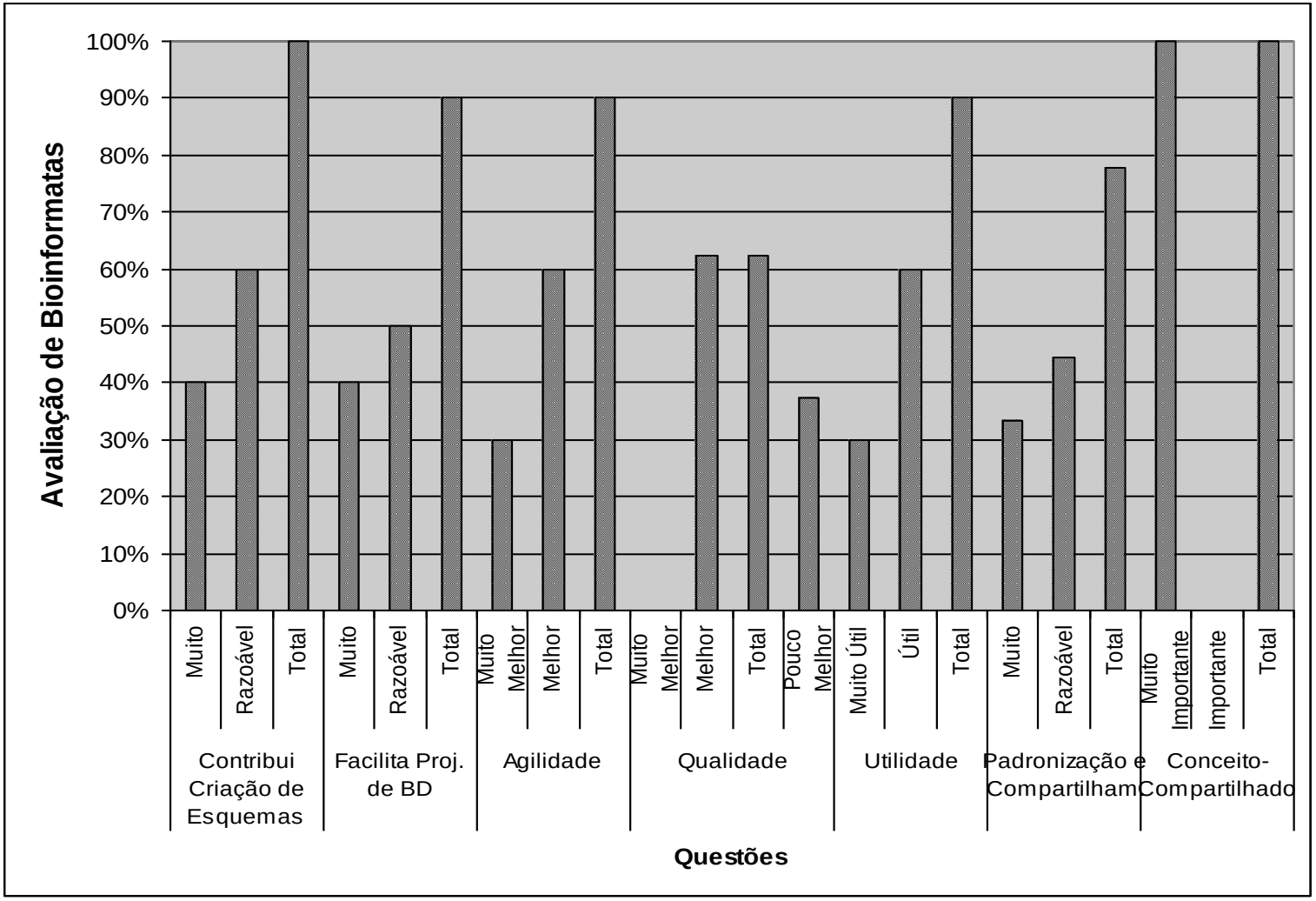
Avaliação dos Usuários



Avaliação dos Usuários



Avaliação dos Usuários



Avaliação dos Usuários