

## PERSISTÊNCIA DE DADOS CLÍNICOS BASEADOS NA MODELAGEM MULTINÍVEL: UMA REVISÃO

Beatriz Proto Martins<sup>1</sup>, Fábio Nogueira de Lucena<sup>1</sup> e Plínio de Sá Leitão Júnior<sup>1</sup>

<sup>1</sup>Instituto de Informática (INF), Universidade Federal de Goiás (UFG), Goiânia/Goiás, Brasil

**Resumo:** Sistemas de Informação em Saúde persistem Registros Eletrônicos de Saúde a partir da definição de como os dados devem ser armazenados e recuperados. Devido ao desenvolvimento de estratégias que empregam a separação entre informação e conhecimento (modelagem multinível), faz-se necessário explorar pesquisas relacionadas à persistência de dados. Objetivos: Investigar o estado da arte referente ao armazenamento e recuperação de dados clínicos baseados na modelagem multinível. Métodos: Revisão sistemática da literatura, envolvendo: requisitos de qualidade para persistência, mapeamento de modelos de banco de dados, descrição e avaliação da camada de persistência e detalhamento de *benchmarks*. Resultados: De 843 artigos encontrados, 45 foram incluídos, mas apenas 20 apresentaram dados relevantes. Conclusão: Em geral, o desenvolvimento de sistemas de saúde multiníveis é orientado por requisitos de desempenho e espaço, contudo as avaliações são centradas no desempenho de consultas. Poucos estudos apresentam estratégias de mapeamento de banco de dados e *benchmarks* significantes.

**Palavras-chave:** Armazenamento e Recuperação da Informação; Sistemas de Informação em Saúde; Registros Eletrônicos de Saúde.

**Abstract:** *Health Information Systems persist Electronic Health Records according to the definition of how data must be stored and retrieved. Due to the development of strategies that employ the separation between information and knowledge (multilevel modeling), it is necessary to explore researches related to data persistence. Objectives: To investigate the state of the art related to the storage and retrieval of clinical data based on multilevel modeling. Methods: Systematic review of the literature, involving: quality requirements for persistence, database model mapping, description and evaluation of the persistence service and benchmark detailing. Results: From 843 papers found, 45 were included, but only 20 presented relevant data. Conclusion: In general, the development of multilevel health systems is driven by performance and space requirements, however evaluations are focused on queries performance. Few studies present significant benchmarks and strategies for database mapping.*

**Keywords:** *Information Storage and Retrieval; Health Information Systems; Electronic Health Records.*

### Introdução

O intercâmbio de Registros Eletrônicos em Saúde (RES) requer o desenvolvimento de Sistemas de Informação em Saúde (SIS) interoperáveis. SISs baseadas no openEHR, no HL7/CDA ou na norma ISO/EN 13606 seguem a modelagem de dados multinível<sup>1</sup>. Essa abordagem possibilita a definição dos dados em vários níveis, compreendendo, pelo menos, o nível de informação, representado pelo Modelo de Referência (MR), e o nível de conhecimento, imediatamente acima, representado pelo Modelo de Conhecimento (MC)<sup>2</sup>.

A persistência de dados clínicos é um fator de impacto na eficácia e na eficiência de SIS, bem como na interoperabilidade desejada entre sistemas. Há várias alternativas para se persistir dados modelados conforme o MC, tal como persistir instâncias de entidades do MR, o qual possui uma estrutura

compartilhada pelas definições do MC. Nesse cenário, é pertinente a investigação de técnicas de persistência no contexto de SIS que seguem o modelo multinível. Este trabalho apresenta uma revisão sistemática que tenta responder os seguintes questionamentos:

1. Questão de pesquisa principal: Quais estratégias de persistência de dados têm sido empregadas em SISs baseados na modelagem multinível?
2. Questão de pesquisa secundária: Quais requisitos, estratégias, modelos de banco de dados (BD), *benchmarks* e avaliações de serviços de persistência de dados têm sido empregados no desenvolvimento de SISs multiníveis?

A revisão sistemática consiste, basicamente, em dois passos: seleção de estudos relevantes, escolhidos a partir da leitura do resumo de cada artigo científico recuperado e; extração de dados dos estudos selecionados, pertinentes às questões de pesquisa com suas respectivas análises. Durante a extração de dados espera-se encontrar requisitos, estratégias, modelos de banco de dados (BD), *benchmarks* e avaliações de serviços de persistência de dados, presentes em SISs cuja modelagem de dados segue a abordagem multinível.

## Métodos

Neste trabalho são executados os seguintes passos para a realização da revisão sistemática: 1) Seleciona-se fontes científicas e artigos de controle; 2) Define-se critérios de inclusão (CI) e de exclusão (CE) baseados nas questões de pesquisa; 2) Aplica-se uma nova *string* de busca nas bases, até que todos os artigos de controle sejam cobertos pela *string*; 3) Para cada artigo, anota-se suas contribuições gerais, relativas às questões de pesquisa, e atribui-se um CI ou CE principal baseado em seu resumo; 4) Para cada estudo coberto pelos CIs, extrai-se os dados relevantes; 4.1) Altera-se o critério do estudo, se necessário.

As fontes bibliográficas escolhidas são: IEEE Xplore (disponível em <http://ieeexplore.ieee.org>); SpringerLink (<https://link.springer.com>); ScienceDirect (<http://www.sciencedirect.com>); Scopus (<https://www.scopus.com/>); PubMed Central (<http://www.ncbi.nlm.nih.gov/pmc/>) e PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>). São recuperados artigos de 2006 a 2016 escritos na Língua Inglesa e Portuguesa.

A *string* avançada de busca, aplicada nas fontes, recupera os artigos de controle<sup>1,3-6</sup> e possui variações por fonte, mas, basicamente, com o mesmo significado semântico: “((*storage* OU *persistence* OU *data model* OU *database schema*) E ((*multilevel* OU *two level* OU *dual*) E (*modeling* OU *approach* OU *methodology*)) OU 13606 OU *information model* OU *reference model*) E (*electronic* E (*health* OU *medical* OU *patient*) E *record*)) E NÃO (*owl*)”.

Conforme a Tabela 1, os CIs abordam estudos que incluem determinados tópicos relacionados ao serviço/camada de persistência de dados em SIS multiníveis.

Tabela 1: CIs e CEs utilizados na seleção de estudos.

CI	Descrição
CI-1	Formulação de requisitos funcionais e não funcionais.
CI-2	Mapeamento do modelo de BD conceitual para o modelo lógico ou físico.
CI-3	Descrição das tecnologias empregadas pelo serviço.
CI-4	Descrição do <i>benchmark</i> utilizado na avaliação.
CI-5	Avaliação da execução do serviço.
CE-1	Não coberto pelos CIs.

CE-2	Contribuições abrangidas por outro estudo incluído (refere-se, também, às duplicações de estudos).
CE-3	Inacessível ou não disponível.
CE-4	Fora do escopo com respeito ao seu propósito.

## Resultados

Na primeira fase de seleção foram incluídos 45 estudos de um total de 843. As quantidades de estudos por fonte e por CI ou CE primário são detalhadas na Tabela 2. Já na segunda fase foram selecionados e extraídos os dados de 20 estudos dos 45, descritos resumidamente na Tabela 3. Deve-se notar que um estudo pode se enquadrar secundariamente em um número arbitrário de CIs e que tais duplicidades inviabilizam o cálculo do total de estudos. Contudo, os critérios secundários de cada estudo são considerados na Seção “Discussão”.

Tabela 2: Quantidades de estudos por fonte bibliográfica e por CIs ou CEs.

	IEEE	PMC	PubMed	ScienceDirect	Scopus	Springer	Total critério
CI-1	0	2	0	0	0	1	3
CI-2	1	1	2	1	3	1	9
CI-3	8	3	6	3	6	1	27
CI-4	0	0	1	0	0	0	1
CI-5	1	1	0	1	0	2	5
Incluídos	10	7	9	5	9	5	45
CE-1	40	96	76	140	301	80	733
CE-2	0	0	5	4	30	6	45
CE-3	0	0	0	0	4	0	4
CE-4	1	0	3	2	6	1	13
Excluídos	41	96	84	146	341	87	795
Total fonte	51	103	93	151	350	92	840

Tabela 3: Quantidade de estudos cobertos por cada CI e CE na segunda fase de seleção.

Critério	(1)	(2)	(3)	(4)	(5)	Total
CI	1	1	13	2	3	20
CE	745	46	4	28	-	823

**Critério CI-1** – A Tabela 4 expõe as características dos estudos que fornecem requisitos próprios ao desenvolvimento da camada de persistência em sistemas multiníveis. O projeto dos mecanismos de persistência deve considerar os requisitos funcionais indispensáveis ao sistema. Estes requisitos descrevem as funcionalidades oferecidas pelos sistemas, que são fortemente ligadas ao armazenamento e recuperação dos dados, como é o caso do serviço de criação, leitura, atualização e deleção (CRUD) de dados. Os requisitos funcionais encontrados nos estudos são retratados na coluna A1.

O projeto da camada de persistência também deve considerar o comportamento esperado pelos usuários em relação à execução do serviço de persistência. Isso é mostrado na coluna A2 da Tabela 4 através de requisitos não-funcionais de desempenho. Estes requisitos exprimem a eficiência no tempo de resposta para a manipulação de dados, mais especificamente, para a execução das operações de CRUD.

É importante que no projeto do serviço de persistência dos estudos sejam consideradas as restrições impostas pelos usuários e pelo sistema sobre o uso de recursos computacionais. Assim, são

expostos na coluna A3, da Tabela 4, os requisitos não-funcionais de espaço relacionados à apropriação da memória.

Devido à necessidade de prover sistemas de saúde interoperáveis, o desenvolvimento dos serviços de persistência desses sistemas deve atender a critérios variáveis e fazer uso de tecnologias heterogêneas. Nesse sentido, os estudos da Tabela 4 definem requisitos não-funcionais de adaptabilidade, portabilidade e compatibilidade na coluna A4.

Tabela 4: Estudos que apresentam requisitos diretamente relacionados à persistência de dados, para o desenvolvimento de sistemas de saúde multiníveis (CI-1).

	Requisitos funcionais (A1)	Desempenho (A2)	Espaço (A3)	Outros requisitos não funcionais (A4)
Wang <i>et al.</i> (2015) <sup>3</sup>	-	Mais rápido que soluções XML, Node+Path, EAV (Entity-Attribute-Value) e ORM (Object-Relational Mapping). Eficiente em ambiente real.	-	BD deve ser adaptável a mudanças nos arquétipos e aos requisitos dos RES.
Velte <i>et al.</i> (2012) <sup>4</sup>	Oferecer CRUD de RES.	Em operações centradas no paciente, as consultas devem ser mais rápidas do que inserções	-	BD dinâmico, gerenciar RES de modo transparente e remoto. Usar web services para CRUD de RES (camada de aplicação independente)
Sundvall <i>et al.</i> (2013) <sup>5</sup>	Permitir CRUD distribuído de RES (sem deleção)	Alto desempenho em ambiente real, utilizando índice e caching no servidor e/ou cliente	Caching no disco Escalabilidade vertical ou horizontal com uso de sharding	Flexível suportando diferentes tecnologias e formatos de armazenamento. Sistema RES transparente, gerenciável
Humm e Walsh (2015) <sup>6</sup>	-	Eficiente no processamento de grandes volumes de RES	Eficiente no espaço utilizado com grande volume de RESs	O modelo de dados deve ser geral, flexível, conveniente/transparente (usar abstrações)
Madaan <i>et al.</i> (2013) <sup>7</sup>	Permitir inserção e consulta de RES.	Eficiente para consultas complexas.	Altamente escalável, favorecido pelo uso de índice.	Adaptação às consultas, sem exigir do usuário conhecimento da estrutura do BD e de sintaxes.

Lianas (2014) <sup>8</sup>	Oferecer CRUD de dados; (de)serializar e criar estruturas de dados.	-	Escalável vertical e horizontal sobre o volume de dados	Assegurar o acesso uniforme e transparente a múltiplos BDs
----------------------------	---	---	---	--

**Critério CI-2** – As propriedades dos estudos que exploram modelos de dados persistentes são mostradas na Tabela 5. O atributo B1 cita o MR do estudo, já a representação do modelo de dados é citada em B2. Detalhes sobre o índice utilizado é apresentado em B3.

Uma vez que o MR e o MC representam grafos de entidades seguindo a abordagem orientada a objetos (OO), é preciso que essas entidades sejam mapeadas para o esquema de BD desejado. Esse mapeamento também é explicado na Tabela 5. Para tanto são relatados os tratamentos dos tipos básicos de dados em B4, de associações e agregações em B5, de herança em B6, de coleções de dados em B7 e mapeamentos específicos ao modelo em B8.

Tabela 5: Estudos que descrevem o mapeamento de modelos conceituais de saúde multiníveis para modelos de dados lógicos ou físicos (CI-2).

	Paul e Hoque (2010) <sup>9</sup>	Wang <i>et al.</i> (2015) <sup>3</sup>
Modelo conceitual (B1)	HL7 RIM	Arquétipos
Modelo lógico ou físico (B2)	EAV-Otimizado (OEAV) sobre a classe Observation e demais	Modelo de BD Relacional
Índice (B3)	Árvore B+ sobre entidade de Observation	Sobre dados de consulta e de identificação
Tipos de dados básicos (B4)	Tipos padrões de BD	Coluna de tipo básico SQL
Associação e agregação (B5)	FK em 1:1 e 1:N; Tabela separada em N:N e agregação	Incorpora coluna em 1:1; Usa FK em 1:N
Herança (B6)	Uma tabela por classe (com identificação) e uma coluna por atributo.	Template mapeia os campos do arquétipo especializado para o do generalizado
Coleção (B7)	-	Tabela com FK para id. do arquétipo e coluna de tipo básico
Mapeamentos específicos (B8)	Campo AV contém um código do Atributo concatenado ao Valor	Tabela contém as versões nova e antigas do arquétipo

**Critério CI-3** – A Tabela 6 realça as propriedades dos serviços de persistência de dados multiníveis apresentados pelos estudos e expõe nas respectivas colunas:

- C1: o nome do sistema ou a quantidade de sistemas em análise, em caso de revisão bibliográfica;
- C2: os tipos de bancos de dados e respectivos Sistemas Gerenciadores de Bancos de Dados (SGBD) utilizados ou interfaces para gerenciamento de dados;
- C3: os padrões de especificação da modelagem dos dados;
- C4: os mecanismos de definição do esquema de dados;
- C5: os mecanismos de consulta dos dados;
- C6: os formatos de organização dos arquivos no BD (tabela, documento, etc).

Tabela 6: Estudos que descrevem a camada ou serviço de persistência de dados em sistemas de saúde com modelagem multinível (CI-3).

	Nome ou quantidade (C1)	BD - SGBD (C2)	Padrão de referência (C3)	Definição de esquema RES (C4)	Consulta RES (C5)	Organização dos arquivos (C6)	
	Humm e Walsh (2015) <sup>6</sup>	-	Relacional - SQL Server	HL7 modificado	-	.NET LINQ	Tabelas geradas por uma ferramenta ORM: Entity, Role e Act
	Frade et al. (2013) <sup>10</sup>	16	Relacional, XML, OO com SQL Server, MySQL e outros	openEHR	-	SQL, AQL, XQuery, SQL+XPath e outros	-
	Wang et al. (2015) <sup>3</sup>	-	Relacional - SQL Server	openEHR	SQL	SQL	Tabela de-normalizada por arquétipo
	Atalag et al. (2011) <sup>11</sup>	GastrOS	Relacional - MS Access e SQLite	openEHR	-	-	Tabela relacional com MR serializado em XML
	López-Nores et al. (2012) <sup>12</sup>	iCabiNET	Relacional - MySQL	EN13606	ADL	-	-
	Osorio et al. (2013) <sup>13</sup>	eHealthCom	XML - BaseX	openEHR	-	-	Documento XML por RES
	Barca et al. (2014) <sup>14</sup>	yourEHRM	NoSQL - MongoDB	openEHR/ EN13606	ADL	AQL	Documento JSON com coluna por caminho de arquétipo
	Bahga e Madiseti (2013) <sup>15</sup>	CHISTAR	Não-relacional - HBase	Estende o openEHR e o HL7	API Java	API Java ou HQL	Tabelas com pares chave/valor conforme tipos de dados do MR
	Flemming et al. (2014) <sup>16</sup>	-	Relacional - MySQL com InnoDB	openEHR modificado	-	-	Tabela por classe do MR gerada pelo Hibernate
	Muñoz et al. (2007) <sup>2</sup>	-	Relacional - Interface ODBC	EN13606	ADL	-	Tabela por classe do MR
	Li et al. (2007) <sup>17</sup>	-	XML nativo - Oracle XML DB	HL7/CDA	-	SQL e XQuery	Documento XML por encontro em diretório hierárquico

Kropf et al. (2015) <sup>18</sup>	-	XML - eXistdb	openEHR	ADL	-	Documentos XML com XForms e RESs
Velte et al. (2012) <sup>4</sup>	-	XML nativo - BaseX	openEHR	ADL	AQL	Documento XML por paciente
Sundvall et al. (2013) <sup>5</sup>	LiU EEE	XML - BaseX, eXistdb	openEHR	ADL	AQL (XQuery interno)	Documento XML por RES
Madaan et al. (2013) <sup>7</sup>	-	NoSQL - MongoDB	openEHR	ADL	AQBE	Documento JSON por arquétipo
Lee et al. (2013) <sup>19</sup>	-	NoSQL - SQL Server	HL7/CDA	CDA com conceitos extraídos do BD avaliado	-	Tabela de pares chave-valor generalizados com conceito individual por linha
Lianas et al. (2014) <sup>8</sup>	PyEHR	Múltiplos - Multi-driver	openEHR	ADL	AQL	-

**Critério CI-4** – Na Tabela 7 são detalhados os elementos dos *benchmarks* utilizados nas avaliações das camadas de persistência. Em D1 é informado o nome/local do repositório do Modelo de Conhecimento. As bases de dados caracterizadas expressam as mesmas informações, apesar de seus formatos variarem conforme as restrições impostas pelos seus respectivos BDs.

Os BDs utilizados com modelagem de um nível (convencional) são citados na Tabela 7 na coluna D2, BDs cujos esquemas seguem o Modelo de Conhecimento em D3 e BDs contendo esquemas conforme o Modelo de Informação em D4. A quantidade de RESs que constituem as bases de dados é exibida em D5 e a quantidade de consultas com semânticas diferentes executadas sobre essas bases é apresentada em D6.

Tabela 7: Estudos que descrevem benchmarks para SISs com modelagem multinível (CI-4).

	Repositório do MC (D1)	BDs de um nível (D2)	BDs com esquemas dependentes do MC (D3)	BDs com esquemas independentes do MC (D4)	Quantidade de RESs (D5)	Quantidade de consultas (D6)
Velte et al. (2012) <sup>4</sup>	-	-	XML nativo	-	1 mil, 10 mil e 30 mil	> 4
Flemming et al. (2014) <sup>16</sup>	CKM e NHS	-	-	Relacional	120	-
Wang et al. (2015) <sup>3</sup>	16 arquétipos do CKM e 1 criado.	Relacional	Relacional (Mapeamento Relacional de Arquétipos - ARM)	Node+Path	~30 mil	7
Madaan et al. (2013) <sup>7</sup>	40 arquétipos do CKM	-	NoSQL	-	-	16

Osorio et al. (2013) <sup>13</sup>	CKM	-	-	XML (insere Compositions no registro novo ou existente)	1 mil, 10 mil e 30 mil	-
Lee et al. (2013) <sup>19</sup>	Conceitos extraídos do BD avaliado	Textual (formatado)	-	NoSQL, XML não nativo e nativo.	1 mil, 5 mil, 10 mil e 50 mil	5
Austin et al. (2009) <sup>20</sup>	SynOD (Dicionário de Objetos Synapses)	-	-	-	3.226	-

**Critério CI-5** – A Tabela 8 apresenta dados sobre os estudos que abordam o critério CI-5. Na coluna E1 são informados os tamanhos resultantes das bases de dados abrangidas pelo trabalho. Nas colunas E2 e E3 são mostradas, respectivamente, o tempo médio gasto na execução de consultas baseadas em indivíduo (registro único) e em população (multi-registros ou agregada). A coluna E4 apresenta o tempo gasto durante a inserção de registros. Por fim, na coluna E5, são descritas, resumidamente, as análises dos autores sobre os resultados relativos à persistência de dados. Uma desvantagem da Tabela 8 é que não é possível fazer comparações precisas entre os estudos, já que eles não utilizam as mesmas bases de dados.

Tabela 8: Estudos que avaliam a camada ou serviço de persistência de dados presentes em sistemas de saúde com modelagem multinível (CI-5).

	Tamanho das bases (E1)	Tempo para consultas de indivíduos (E2)	Tempo para consultas de populações (E3)	Tempo inserção (E4)	Análise dos resultados (E5)
Velte et al. (2012) <sup>4</sup>	-	0,65 ms	Mil: 20 ms; 10 mil: 176 ms; 30 mil: 523 ms	Aumenta com o volume. 30 mil: 423 ms	Eficaz para consulta baseada em indivíduos.
Wang et al. (2015) <sup>3</sup>	Relacional: 1,6 GB; ARM: 2,9 GB; No-de+Path: 43,87 GB	Relacional: 272 ms; ARM: 243 ms; No-de+Path: 165.246 ms	Relacional: 6.082 ms; ARM: 379 ms; Node+Path: 10.708 ms	-	Consulta agilizada pelos índices e <i>caching</i> de dados em arquétipos descendentes
Paul e Hoque (2010) <sup>9</sup>	EAV = 1,26*OEAV	EAV = 15,15*OEAV	EAV = 56,3*OEAV	-	Eficiente para buscas, dados esparsos e de alta dimensão; Consome menos espaço que EAV
Muñoz et al. (2007) <sup>2</sup>	-	-	-	-	Não opera em um cenário real
Austin et al. (2009) <sup>20</sup>	-	-	-	-	Alocação explícita de memória gera desperdício



Li et al. (2007) <sup>17</sup>	-	-	10,2 ms	15,08 ms	Índices com Árvore B e XPath melhoram, respectivamente, consultas baseadas em indivíduo e população
Osorio et al. (2013) <sup>13</sup>	-	1 mil: 4 ms; 10 mil: 5 ms; 30 mil: 5 ms	1 mil: 126 ms; 10 mil: 232 ms; 30 mil: 594 ms	1 mil: 34 ms; 10 mil: 36 ms; 30 mil: 49 ms	Persistência de RES é eficiente, ao contrário de consulta baseada em população e inserção
Lee et al. (2013) <sup>19</sup>	Textual: 120 MB	-	[50 mil RESs] Chave-valor generalizado: 1.414 ms; XML não nativo: 4.722 ms; XML nativo: 8.407 ms	-	Chave-valor generalizada é mais rápida, já o tamanho do XML é mínimo. Ambos tem esquema flexível.

## Discussão

**Critério CI-1** – Os estudos recomendam que o modelo de BD escolhido seja flexível, facilitando a adição e remoção de entidades e atributos<sup>3,4,6</sup>, possivelmente livre de esquema<sup>7</sup>, e que ofereça suporte a diferentes tecnologias e formatos de armazenamento<sup>5,8</sup>. Além disso, deve ser assegurado o acesso transparente a múltiplos BDs por meio de *drivers*<sup>8</sup> ou ao BD único através de abstrações para gerenciar os RESs<sup>4,5,6</sup>. Isso inclui não exigir do usuário conhecimento da estrutura do BD (usando o mapeamento objeto-relacional<sup>6,7</sup>) e das sintaxes<sup>7</sup> na realização de consultas.

Os estudos esperam que o sistema seja eficiente no acesso (leitura/escrita) de um grande volume de RESs<sup>6</sup> e aplicável em um ambiente real<sup>3,5</sup>. As operações devem ser centradas no paciente com consultas mais rápidas do que inserções<sup>4</sup>. Nesse sentido, a camada de persistência deve permitir a execução de consultas complexas<sup>7</sup>. Isso pode ser alcançado através de *caching* no servidor e/ou no cliente<sup>5</sup> e através de índices<sup>5,7</sup> que otimizam a consulta<sup>7</sup>.

Além do desempenho, o espaço é outro requisito relevante. O sistema deve propiciar escalabilidade horizontal e/ou vertical sobre o volume de dados<sup>5,8</sup>. A escalabilidade pode ser alcançada através do suporte a múltiplos sistemas de armazenamento<sup>8</sup> ou pelo uso de *sharding*<sup>5</sup> (divisão do BD). O uso de índices<sup>7</sup> e de armazenamento de *caching* em disco<sup>5</sup> também são úteis.

Em síntese, os requisitos dos artigos para o desenvolvimento da camada de persistência centram-se em critérios de flexibilidade e adaptabilidade do modelo de BD e de desempenho na manipulação dos dados. Esses requisitos devem ser atingidos considerando a redução do tempo de resposta para consultas.

**Critério CI-2** – Foram encontradas duas estratégias que detalham o mapeamento de instâncias do modelo de BD conceitual para o modelo de BD lógico, são elas: mapeamento do HL7 RIM (MR do HL7/CDA) para o modelo relacional com a aplicação do modelo Entidade-Atributo-Valor-Otimizado (OEAV) sobre a classe Observation<sup>9</sup> e; mapeamento de arquétipos para o modelo relacional<sup>3</sup>.

As estratégias desenvolvidas evitam a perda de desempenho e de espaço na camada de persistência. No caso dos arquétipos, diferentes versões de um mesmo arquétipo são organizadas em uma única tabela<sup>3</sup>. No caso do HL7 RIM, uma variável Atributo-Valor de tipo inteiro armazena a representação binária concatenada do atributo e do valor da classe Observation no modelo OEAV<sup>9</sup>.

Os tipos de dados básicos do modelo conceitual, com no máximo uma ocorrência, são mapeados para os tipos padrões de BD<sup>3,9</sup>. A classe *DvBoolean* com seu atributo *value*, por exemplo, é mapeada para uma coluna SQL de tipo *INTEGER*<sup>3</sup>. A chave primária em cada tabela relacional corresponde ao item de identificação do arquétipo instanciado ou a um valor gerado<sup>3</sup> (chave substituta). No caso da tabela *OEAV*, a chave é baseada na identificação do paciente<sup>9</sup>.

Tanto itens de dados em arquétipos sem limites de ocorrências, quanto tipos de dados de coleção no HL7 RIM são mapeados para tabelas com duas colunas: uma coluna com uma chave estrangeira (FK) referenciando o arquétipo, ou classe ao qual o item pertence, e outra coluna com os dados propriamente ditos<sup>3,9</sup>. Para isso é necessário criar uma tabela para cada atributo do tipo coleção<sup>9</sup>.

Associações entre tabelas, representando arquétipos ou classes do HL7 RIM, do tipo um-para-muitos (i.e. atual-para-alvo) são mapeadas como FK na tabela alvo para a tabela atual<sup>3</sup> ou vice-versa<sup>9</sup>. Em associações um-para-um, a tabela alvo é concatenada na tabela atual<sup>3</sup> ou utiliza-se uma FK<sup>9</sup>. No relacionamento muitos-para-muitos, é criada uma nova tabela com FKs para as duas tabelas associadas.

Em relação à herança, o mapeamento fica a cargo de *templates* que associam os campos de arquétipos especializados a um subconjunto de campos correspondentes nos arquétipos generalizados<sup>3</sup>. Para instâncias HL7 RIM, é criada uma tabela por classe, de modo a não armazenar atributos herdados<sup>9</sup>.

Uma forma de otimizar a persistência é através do uso de índices. Para tanto, os estudos recomendam o uso de índices nos itens de dados de consulta<sup>3</sup> e de identificação<sup>3,9</sup>, este último relacionado à Entidade *OEAV* da classe *Observation*.

Ambas as estratégias<sup>3,9</sup> fazem a conversão dos modelos de BDs com técnicas de mapeamento amplamente conhecidas na área de Banco de Dados, como no caso do tratamento de herança e associação de entidades. Isto significa que são aplicadas estratégias comuns, em sua maioria, no cenário específico de dados em saúde. Além disso, o BD relacional baseado em arquétipos<sup>3</sup> pode ser inflexível, já que exige que todos os conceitos sejam previamente conhecidos.

**Critério CI-3** – Com exceção de um artigo de revisão<sup>10</sup>, os estudos que apresentam a camada ou serviço de persistência de dados multinível, abordam um único sistema ou arquitetura. Para tanto, são relatadas as tecnologias usadas na persistência e o formato de organização dos arquivos no BD. Dentre os estudos são descritos os sistemas *GastrOS*<sup>10,11</sup>, *iCabiNET*<sup>12</sup>, *eHealthCom*<sup>13</sup>, *LiU EEE*<sup>5,10</sup>, *yourEHRM*<sup>14</sup>, *CHISTAR*<sup>15</sup> e *pyEHR*<sup>8</sup>.

Os estudos utilizam os seguintes BDs e Sistemas Gerenciadores de Bancos de Dados (SGBDs): BDs relacionais com os SGBDs *SQL Server*<sup>3,6,10</sup>, *MS Access*<sup>11</sup>, *SQLite*<sup>10,11</sup>, *MySQL*<sup>10,12,16</sup> ou fornecem uma interface *ODBC*<sup>2</sup>; BDs XML com os SGBDs *Oracle XML DB*<sup>10,17</sup>, *BaseX*<sup>1,4,5,10,13</sup>, *eXistdb*<sup>1,5,10,18</sup> e *Berkeley DB XML*<sup>1</sup>; BDs NoSQL com os SGBDs *CouchBase*<sup>1</sup>, *MondoDB*<sup>7,14</sup> e *SQL Server*<sup>19</sup>; BD não-relacional com o SGBD *HBase*<sup>15</sup> e BDs múltiplos<sup>8</sup>, isto é, diferentes tipos de BDs suportados pelo serviço, desde que os sistemas que o utilizem implementem a interface *multi-driver* fornecida.

Nos estudos são usados os seguintes padrões de definição multinível da estrutura dos dados a serem persistidos: *openEHR*<sup>1,3,5,7,8,10,11,13,14,18</sup>, *EN13606*<sup>2,12,14</sup>, *HL7/CDA*<sup>17,19</sup>, *HL7 modificado*<sup>6,15</sup> e *openEHR modificado*<sup>15,16</sup>. Com base nesses padrões, são utilizados os seguintes mecanismos para consulta dos esquemas de dados: *AQL*<sup>4,5,8,10,14</sup>, *SQL*<sup>3,10,17</sup>, *XQuery*<sup>5,10,17</sup>, *.NET LINQ*<sup>6</sup>, *AQBE*<sup>7</sup>, *HQL*<sup>15</sup> e *API Java*<sup>15</sup>.

Os estudos do critério CI3 geralmente descrevem, de forma sucinta, a estrutura do modelo de BD utilizado para persistência. Alguns BDs de SISs multiníveis são formados por tabelas relacionais, em que cada tabela representa um arquétipo<sup>3</sup> ou uma classe do MR<sup>2,6,16</sup>. Essas tabelas são geradas pelo mapeamento manual<sup>3</sup> e implementadas com o apoio de ferramentas que auxiliam no mapeamento objeto-relacional<sup>6,16</sup>, por exemplo, pelo *Hibernate*<sup>16</sup>. Também é possível serializar objetos do MR de/para o formato XML e persisti-los em tabelas relacionais<sup>11</sup>.

Na maioria dos casos o BD é composto por documentos XML, seja por paciente<sup>1,4</sup> ou por RES<sup>5,13,18</sup> ou, ainda, por visita em um diretório hierárquico<sup>17</sup>. BDs NoSQL são constituídos por tabelas de pares

chave-valor generalizados<sup>19</sup> (contendo conceito por linha) ou por documentos JSON. Estes documentos JSON são gerados por paciente<sup>1</sup> ou por arquétipo, contendo uma coluna para cada caminho do arquétipo<sup>7,10</sup>. No caso do BD não-relacional, são armazenadas tabelas com pares chave-valor conforme os tipos de dados do MR<sup>15</sup>.

**Critério CI-4** – Os *benchmarks* dos estudos são formados por bases de dados conforme o modelo de um nível (convencional), das quais, cinco, no total, são relacionais<sup>1,3</sup> e uma é estruturada textualmente<sup>19</sup>. Em relação aos dados modelados com a abordagem multinível, os *benchmarks* são compostos por bases relacionais<sup>1,3,16</sup>, XML<sup>1,4,13,19</sup>, NoSQL<sup>1,7,19</sup> e Node+Path<sup>3</sup>.

Para os BDs baseados na modelagem multinível, foram utilizados os MCs presentes nos repositórios CKM<sup>1,3,7,13,16</sup>, NHS<sup>16</sup> e repositórios próprios não divulgados<sup>1,3,19,20</sup>, incluindo um repositório de SynOD (Dicionário de Objetos Synapses)<sup>20</sup> e um com conceitos extraídos do BD empregado na avaliação<sup>19</sup>. Em alguns casos<sup>1,3,7</sup>, apenas parte do repositório é importado para o *benchmark*.

As bases de dados dos *benchmarks* são compostas por RESs, com as seguintes quantidades máximas por estudo: 120<sup>16</sup>, 3.226<sup>20</sup>, 30 mil<sup>3,4,13</sup>, 50 mil<sup>19</sup> e 4,2 milhões<sup>1</sup>. Destes registros apenas um conjunto é sintético<sup>13</sup> (gerado com base em amostras). Em média, são disponibilizadas 8 consultas por estudo<sup>1,3,4,7,19</sup>. Apesar de serem fornecidos *benchmarks* com dados reais, a quantidade de registros é pequena, exceto em um estudo<sup>1</sup>, se comparada à de sistemas reais.

**Critério CI-5** – O modelo relacional<sup>1,3</sup> é o modelo de BD baseado na abordagem multinível que menos consome espaço, quando comparado ao Node+Path, ARM (Archetype Relational Mapping)<sup>3</sup>, NoSQL e XML<sup>1</sup>. Consultas baseadas em indivíduos são mais rápidas no BD Node+Path do que nas soluções relacional e ARM<sup>3</sup>. Por outro lado, não são recomendadas consultas baseadas em população em BDs XML<sup>1,19</sup>.

Consultas baseadas em indivíduos são mais rápidas do que consultas baseadas em população<sup>3,4,13</sup> e do que inserções de Compositions em RESs<sup>13</sup>. Além disso, as consultas podem ser agilizadas pelo uso de índices<sup>1,3,17</sup> e *caching*<sup>3</sup>. A solução chave-valor generalizada é rápida, mas tem um tamanho significativo<sup>19</sup>. Adicionalmente, a alocação de coleções deve ser dinâmica para evitar desperdício de memória<sup>20</sup>. Com relação à eficiência dos modelos de dados, o OEAV é mais eficiente do que o EAV<sup>9</sup>. Por outro lado, uma das soluções relacionais apresentadas não consegue operar em cenário real<sup>2</sup>.

## Conclusão

Em resposta à questão de pesquisa principal, foram encontradas estratégias de persistência que realizam o mapeamento do MR e do MC para BDs amplamente conhecidos. Para isso, são utilizadas técnicas de mapeamento comuns e não técnicas otimizadas que contemplem as especificidades do modelo multinível.

Em resposta à questão secundária, os estudos requerem, geralmente, o uso de um BD que seja flexível na definição do esquema e eficiente em desempenho e espaço. As soluções dos estudos mostram-se mais vantajosas para consultas baseadas em indivíduo do que para qualquer outra operação sobre o BD. Um dos *benchmarks* encontrados apresenta um volume de dados significativamente superior aos demais.

Como trabalho futuro, espera-se desenvolver uma estratégia de persistência que atenda os requisitos de qualidade (não funcionais) identificados e que considere as especificidades do modelo multinível. O objetivo é mapear o MR para uma estrutura de dados altamente otimizada e compará-la às estratégias encontradas.

## Agradecimentos

Este trabalho foi realizado com o apoio financeiro do Programa de Demanda Social fornecido pela CAPES, fundação do Ministério da Educação do Brasil, à autora Beatriz Proto Martins.

## Referências

- [1] Freire SM, Teodoro D, Wei-Kleiner F, Sundvall E, Karlsson D, Lambrix P. Comparing the Performance of NoSQL Approaches for Managing Archetype-Based Electronic Health Record Data. *PLoS One*. 2016 Mar 9;11(3):e0150069.
- [2] Muñoz A, Somolinos R, Pascual M, Fragua JA, González MA, Monteagudo JL, et al. Proof-of-concept design and development of an EN13606-based electronic health care record service. *J Am Med Inform Assoc*. 2007. Jan-Feb;14(1):118-129. Epub 2006 Oct 26. PubMed PMID: 17068357.
- [3] Wang L, Min L, Wang R, Lu X, Duan H. Archetype relational mapping - a practical openEHR persistence solution. *BMC Med Inform Decis Mak*. 2015 Nov 5;15:88. doi: 10.1186/s12911-015-0212-0. Erratum in: *BMC Med Inform Decis Mak*. 2016;16:21. PubMed PMID: 26541142; PubMed Central PMCID: PMC4636072.
- [4] Velte L, Pedrosa T, Costa C, Oliveira J. An openEHR repository based on a native XML database. *Proceedings of the International Conference on Health Informatics*. ISBN 978-989-8425-88-1. 2012;386-389.
- [5] Sundvall E, Nyström M, Karlsson D, Eneling M, Chen R, Öрман H. Applying representational state transfer (REST) architecture to archetype-based electronic health record systems. *BMC Med Inform Decis Mak*. 2013 May 9;13:57. doi: 10.1186/1472-6947-13-57. PubMed PMID: 23656624.
- [6] Humm G, Walsh P. Flexible yet Efficient Management of Electronic Health Records. *International Conference on Computational Science and Computational Intelligence*. 2015;771-775. doi: 10.1109/CSCI.2015.84.
- [7] Madaan A, Chu W, Daigo Y, Bhalla S. Quasi-Relational Query Language Interface for Persistent Standardized EHRs: Using NoSQL Databases. *Databases, Networked Information Systems. DNIS 2013, LNCS 7813*. 2013;182-196.
- [8] Lianas L, Frexia F, Delussu G, Anedda P, Zanetti G. pyEHR: A scalable clinical data management toolkit for biomedical research projects. *IEEE 16th International Conference on e-Health Networking, Applications and Services, Natal*. 2014;370-374. DOI: 10.1109/HealthCom.2014.7001871.
- [9] Paul R, Hoque ASML. Search efficient representation of healthcare data based on the HL7 RIM. *Journal of Computers*. 2010;5(12):1810-1818. DOI: 10.4304/jcp.5.12.1810-1818
- [10] Frade S, Freire SM, Sundvall E, Patriarca-Almeida JH, Cruz-Correia R. Survey of openEHR storage implementations. *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, Porto*. 2013;303-307. doi: 10.1109/CBMS.2013.6627806
- [11] Atalag K, Yang HY, Tempero E, Warren J. Model driven development of clinical information systems using openEHR. *Stud Health Technol Inform*. 2011;169:849-53. PubMed PMID: 21893867.
- [12] López-Nores M, Blanco-Fernández Y, Pazos-Arias JJ, García-Duque J. The iCabiNET system: Harnessing Electronic Health Record standards from domestic and mobile devices to support better medication adherence. *Computer Standards & Interfaces*. Jan 2012;34(1):109-116. ISSN 0920-5489. DOI: <http://dx.doi.org/10.1016/j.csi.2011.05.012>.
- [13] Osorio E, Ferreira L, Abreu R, Sousa F. Interoperability in Ambient Assisted Living using OpenEHR. *IEEE 15th International Conference on e-Health Networking, Applications & Services*. 2013;394-398. Doi: 10.1109/HealthCom.2013.6720707

- [14] Barca CC, Lagunar CM, Rodríguez JM, Quintero AM, Martins IRM, Martínez I, et al. yourEHRM: Standard-based management of your personal healthcare information. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Valencia. 2014;89-92. Doi: 10.1109/BHI.2014.6864311
- [15] Bahga A, Madiseti VK. A cloud-based approach for interoperable electronic health records (EHRs). IEEE J Biomed Health Inform. 2013 Sep;17(5):894-906. doi: 10.1109/JBHI.2013.2257818. PubMed PMID: 25055368.
- [16] Flemming D, Paul M, Hübner U. Building a common ground on the clinical case: design, implementation and evaluation of an information model for a Handover EHR. Stud Health Technol Inform. 2014;201:167-74. PubMed PMID: 24943540.
- [17] Li H, Duan H, Lu X, Huang Z. A clinical document repository for CDA documents. 1st International Conference on Bioinformatics and Biomedical Engineering, ICBBE, 4272764. 2007;1084-1087. DOI: 10.1109/ICBBE.2007.280.
- [18] Kropf S, Chalopin C, Denecke K. Template and Model Driven Development of Standardized Electronic Health Records. Stud Health Technol Inform. 2015;216:30-4. PubMed PMID: 26262004.
- [19] Lee KKY, Tang WC, Choi KS. Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage. Computer Methods and Programs in Biomedicine. 2013; April 10(1):99-109, ISSN 0169-2607. DOI: <http://dx.doi.org/10.1016/j.cmpb.2012.10.018>.
- [20] Austin T, Kalra D, Lea NC, Patterson DL, Ingram D. Analysis of Clinical Record Data for Anticoagulation Management within an EHR System. Open Med Inform J. 2009 Aug 19;3:54-64. doi: 10.2174/1874431100903010054. PubMed PMID: 19738915; PubMed Central PMCID: PMC2737130.

## Contato

Beatriz Proto Martins.  
Estudante de mestrado do Instituto de  
Informática (INF/UFG).  
Alameda Palmeiras, Quadra D, Câmpus  
Samambaia. CEP 74690-900 – Goiânia - GO.  
Fone: (62)3521-1181 / Fax: (62)3521-1182  
[beatrizprotomartins@inf.ufg.br](mailto:beatrizprotomartins@inf.ufg.br)

