Extração, Integração e Importação de Dados Heterogêneos em Cidades Inteligentes: Um Mapeamento Sistemático

João Gabriel Almeida, Jorge Silva, Everton Cavalcante, Thais Batista

Universidade Federal do Rio Grande do Norte (UFRN)
Natal-RN, Brasil

j.quaresmasantos_98@hotmail.com, jorgepereirasb@gmail.com
everton@dimap.ufrn.br, thaisbatista@gmail.com

Resumo. *Dados exercem um papel fundamental em cidades inteligentes, sendo utilizados por usuários e aplicações para os mais diferentes propósitos. Como consequência da diversidade de dispositivos e sistemas existentes e da falta de padronização em termos de modelos, protocolos e formatos de dados adotados, observa-se que os dados de cidades inteligentes sãso bastante heterogêneos. Tal heterogeneidade impõe dificuldades ao desenvolvimento de aplicações e utilização desses dados, fazendo com que seja necessário investigar estratégias e soluções para extração, integração e importação de dados em ambientes de cidades inteligentes. A fim de prover um panorama do atual estado da arte com relação a esse tópico, foi realizado um mapeamento sistemático no qual 28 estudos disponíveis na literatura foram coletados e analisados para identificar estratégias, metodologias e facilidades existentes para apoiar e implementar essas atividades. Este artigo apresenta os principais resultados do mapeamento sistemático realizado, bem como um conjunto de desafios de pesquisa e desenvolvimento que podem direcionar trabalhos futuros.*

Abstract. *Data play a key role in smart cities, being used by users and applications for many purposes. As a consequence of the diversity of existing devices and systems and the lack of standardization in terms of the adopted data models, protocols, and formats, smart city data are significantly heterogeneous. Such a heterogeneity hampers application development and use of these data, thus requiring the investigation of strategies and solutions to extract, integrate, and import data in smart city environments. Aiming at providing a panorama of the current state of the art on this topic, a systematic mapping was carried out with 28 studies, which were collected and analyzed to identify existing strategies, methodologies, and facilities to support and implement these activities. This paper presents the main results of the conducted systematic mapping and a set of research and development challenges that may drive future work.*

1. Introdução

Uma das principais características de cidades inteligentes diz respeito à miríade de dados gerados por uma grande diversidade de dispositivos e sistemas existentes, os quais têm sido frequentemente desenvolvidos e implantados em significativa fragmentação e isolamento, cada um sendo responsável por lidar com áreas, preocupações e problemas específicos [Souza et al. 2017]. Além disso, observa-se nesse contexto uma alta heterogeneidade de dados, em termos de padrões, formatos e protocolos adotados, contribuindo para a existência de silos verticais de dados [d'Aquin et al. 2015].

A inerente heterogeneidade de dados nesse cenário acaba impondo desafios à plena concretização do conceito de cidades inteligentes, uma vez que dados são fundamentais para apoiar processos de tomada de decisão. Como discutem [von Landesberger et al. 2016], essas questões dificultam a utilização das informações disponíveis da cidade e o desenvolvimento de aplicações que possam permitir uma gestão mais eficiente dos problemas enfrentados por ela. Para que esses problemas sejam superados, é imprescindível a proposição e implementação de estratégias e soluções para extração, integração e importação de dados em ambientes de cidades inteligentes, no intuito de normalizá-los, padronizá-los, facilitar a troca de informações e garantir maior integridade e consistência das informações geradas.

Embora a integração de dados heterogêneos seja importante para a concretização de sistemas para cidades inteligentes, ainda existem lacunas no que se refere ao entendimento de que estratégias poderiam ser adotadas nesse cenário. Recentemente, [Silva et al. 2018] realizaram uma revisão de literatura que analisou onze estudos publicados entre 2015 e 2017 a fim de identificar as principais estratégias utilizadas no desenvolvimento de soluções de integração, relacionamento e representação de dados em cidades inteligentes. Essa revisão de literatura, apesar de ter empregado um processo sistemático, considerou um intervalo de tempo específico e não considerou, para além da integração, questões relacionadas à extração e importação de dados heterogêneos. Além disso, esse estudo considerou apenas três bases eletrônicas de publicação.

Este artigo apresenta os resultados de um mapeamento sistemático realizado com o objetivo de prover uma visão geral atualizada da literatura acerca de soluções para extração, integração e importação de dados heterogêneos em cidades inteligentes. Um mapeamento sistemático é um tipo de estudo secundário que visa, através de um procedimento sistemático rigoroso de coleta, seleção e análise de estudos disponíveis na literatura, obter uma visão abrangente acerca de um determinado tópico, identificar lacunas em pesquisa e desenvolvimento, bem como e coletar evidências que podem dar subsídios para a condução de outros estudos [Kitchenham and Charters 2007, Petersen et al. 2008]. No mapeamento sistemático realizado, 28 estudos foram selecionados e analisados no tocante à proposição de estratégias, soluções e facilidades para viabilizar a integração, extração e importação de dados heterogêneos. Além disso, o estudo também apresenta alguns desafios relevantes de pesquisa e desenvolvimento relacionados a essas questões.

O restante desse artigo está estruturado da seguinte forma. A Seção 2 apresenta a metodologia adotada neste trabalho em termos das questões de pesquisa a serem respondidas e as estratégias de busca e seleção dos estudos. A Seção 3 provê uma síntese resultante da análise dos estudos como respostas às questões de pesquisa estabelecidas. A Seção 4 aponta algumas questões importantes que podem direcionar pesquisas futuras. Por fim, a Seção 5 sumariza os principais resultados deste estudo.

2. Metodologia

O mapeamento sistemático apresentado neste artigo foi conduzido seguindo orientações disponíveis na literatura para a realização desse tipo de estudo. As etapas básicas de um mapeamento sistemático são: (i) *planejamento*, etapa que resulta em um protocolo definindo as questões de pesquisa a serem investigadas, a estratégia de busca a ser adotada, os critérios a serem utilizados para selecionar estudos e os métodos para extração e síntese de dados; (ii) *execução*, na qual estudos são identificados, selecionados e avaliados

de acordo com o protocolo definido, e; (iii) *relato*, que agrega informações extraídas dos estudos relevantes considerando as questões de pesquisa e estabelece conclusões.

Questões de pesquisa. No intuito de encontrar estudos na literatura que apresentem estratégias e soluções para facilitar os processos de extração, integração e importação de dados no contexto de cidades inteligentes, as seguintes questões de pesquisa (QPs) foram definidas:

QP1: Quais as estratégias, metodologias e ferramentas atualmente utilizadas para a prover a integração, extração ou importação de dados heterogêneos em ambientes de cidades inteligentes?

QP2: Quais as facilidades oferecidas pelas abordagens existentes para integração, extração ou importação de dados provenientes de fontes heterogêneas?

Na QP1, as estratégias podem ser definidas como ferramentas, *frameworks*, bibliotecas ou mesmo aplicações que tratem da extração, integração ou importação de dados oriundos de fontes heterogêneas no contexto de cidades inteligentes. Por sua vez, a QP2 aprofunda a questão de como tais estratégias facilitam os processos anteriormente citados, a exemplo de uma interface que interage com o usuário de maneira coesa ou uma API bem definida voltada a desenvolvedores de aplicações.

Estratégia de busca. Para recuperar estudos a partir da literatura, foi utilizado um processo automatizado de busca realizado sobre cinco bases eletrônicas de publicação, IEEEXplore, ACM Digital Library, ScienceDirect.com, Scopus e Web of Knowledge. A escolha por essas bases de publicação levou em consideração o fato de elas estarem entre as bases mais populares na área de Computação e outros critérios importantes, tais como cobertura da literatura, qualidade dos resultados retornados pelo procedimento de busca automática, disponibilidade de texto completo dos estudos, facilidade de uso, regularidade de atualização e versatilidade na exportação de resultados [Zhang and Babar 2010].

Com base nas QPs definidas, quatro termos principais foram inicialmente identificados, a saber, *integração de dados*, *extração de dados*, *importação de dados* e *cidades inteligentes*. Esses termos constituíram a seguinte *string* de busca (em Inglês):

```
(data integration OR data extraction OR data import) AND
                (smart city OR smart cities)
```

Critérios de seleção. Foram utilizados critérios para avaliar cada estudo de acordo com as QPs definidas com o objetivo de incluir estudos potencialmente relevantes para respondê-las e excluir aqueles que não contribuiriam para respondê-las.

Dois critérios de inclusão (CIs) foram considerados:

CI1: O estudo apresenta soluções/ferramentas para a integração, extração ou importação de dados heterogêneos em ambientes de cidades inteligentes.

CI2: O estudo apresenta facilidades para a integração, extração ou importação de dados heterogêneos em ambientes de cidades inteligentes.

Cinco critérios de exclusão (CEs) foram também definidos:

CE1: O estudo não apresenta e/ou discute sobre soluções para a integração, extração ou importação de dados heterogêneos em ambientes de cidades inteligentes.

CE2: O estudo é uma versão anterior de um estudo mais completo sobre a mesma pesquisa.

CE3: O estudo não possui um resumo ou o texto completo não está disponível.
CE4: O estudo é um índice, prefácio, tutorial, editorial, palestra ou resumo de conferência/workshop.
CE5: O texto do estudo não está escrito em Inglês, que é o idioma mais comum em materiais acadêmicos.

Para este mapeamento sistemático, um estudo foi considerado relevante se ele tiver satisfeito pelo menos um CI e não tiver satisfeito qualquer CE.

Extração e síntese de dados. Para extrair dados dos estudos selecionados e sintetizar os resultados, uma planilha de extração de dados foi organizada com itens relacionados às QPs e outras informações relevantes. Além de informações básicas como título do estudo, ano e veículo de publicação, os dados extraídos versaram sobre: (i) a estratégia utilizada para gerenciar os diferentes dados coletados, através do desenvolvimento de um *framework*, biblioteca, sistema Web, aplicação móvel, etc.; (ii) o domínio de aplicação no qual cada solução apresentada pelos estudos é empregada; (iii) a abordagem utilizada para facilitar a extração, integração ou importação de dados heterogêneos; (iv) as fontes de dados utilizadas para extração de dados, tais como APIs, bancos de dados relacionais e não-relacionais, arquivos, etc.; (v) o tipo de modelo de dados utilizado em cada estudo para padronizar os dados extraídos, e; (vii) a estratégia usada para armazenamento de dados, como o uso de bancos de dados relacionais e não-relacionais, sistemas de arquivos distribuídos, uso de serviços de nuvem, etc.

Processo de seleção. As atividades envolvidas no mapeamento sistemático foram conduzidas entre os meses de agosto e outubro de 2019, de modo que foram considerados estudos publicados até esse período. Durante o processo de busca, a *string* de busca foi adaptada para torná-la compatível com o mecanismo de busca de cada base eletrônica de publicação. Em seguida, o procedimento automatizado de busca foi realizado sobre cada base utilizando a *string* adaptada. A busca limitou-se às informações de título, resumo e palavras-chave de cada estudo retornado.

Após recuperar os estudos a partir das bases eletrônicas de publicação, o processo de seleção foi conduzido em quatro etapas. Na primeira etapa, os estudos retornados pelo procedimento automatizado de busca foram unificados em um repositório único e duplicatas foram removidas. A segunda etapa envolveu a leitura de título, resumo e palavras-chave dos estudos recuperados, os quais foram filtrados de acordo com os critérios de seleção definidos (CIs/CEs). A terceira etapa envolveu a leitura das seções de introdução e conclusão dos estudos e uma nova aplicação dos critérios de seleção. Por fim, a quarta etapa consistiu da leitura completa dos estudos e preenchimento da planilha de extração de dados. A Figura 1 mostra a execução dessas etapas, a qual resultou em um conjunto de 28 estudos selecionados como relevantes, identificados como E1 a E28 no Apêndice A.

3. Resultados

Esta seção sumariza os resultados do mapeamento sistemático realizado considerando as QPs e os dados extraídos/sintetizados a partir dos estudos analisados. A Seção 3.1 traz uma breve visão geral dos estudos selecionados, enquanto as Seções 3.2 e 3.3 apresentam as respostas a cada QP.

3.1. Visão Geral dos Estudos Selecionados

Distribuição ao longo dos anos. A Figura 2 mostra a distribuição das publicações nos últimos oito anos (2012-2019). Pelo gráfico apresentado, é possível observar um relevante

Figura 1. Etapas para seleção dos estudos relevantes.

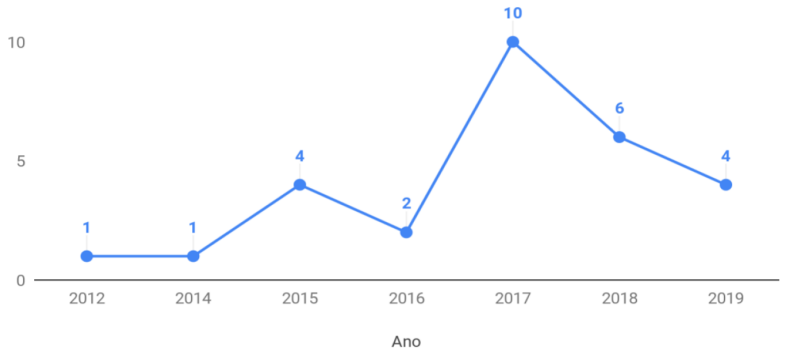

Figura 2. Publicação dos estudos selecionados ao longo dos anos.

crescimento no número de estudos sobre manipulação de dados heterogêneos ao longo dos anos, com uma média de 3,5 estudos publicados por ano.

Domínios de aplicação. Com relação aos domínios de aplicação em que as soluções apresentadas pelos estudos primários selecionados se inserem, pode-se observar que o foco da maioria das soluções propostas pelos estudos relevantes selecionados é abranger, de maneira geral, diversos domínios relacionados ao contexto de cidades inteligentes, tais como meio ambiente, mobilidade, dentre outros. A Figura 3 mostra os diferentes domínios de aplicação das abordagens para integração de dados heterogêneos em cidades inteligentes identificadas nos estudos selecionados.

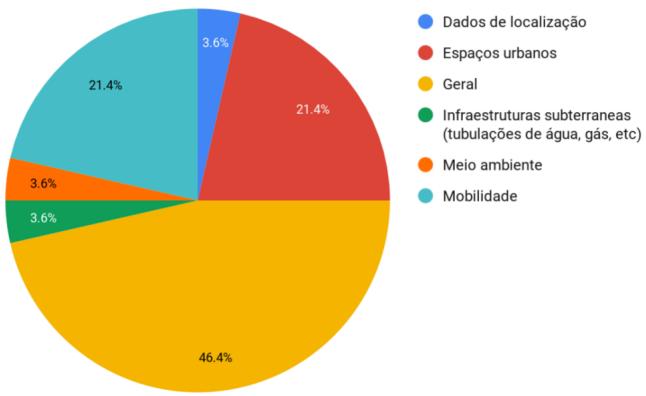

Figura 3. Domínios de aplicação das soluções propostas pelos estudos relevantes selecionados.

3.2. Estratégias para Integração, Extração e Importação de Dados

Abordagens para integração de dados. Com base nos estudos analisados, quatro abordagens principais foram identificadas para integração de dados em ambientes de cidades inteligentes, a saber: (i) *consolidação de dados*, (ii) *federação de dados*, (iii) *propagação de dados* e (iv) *vocabulário controlado*. A Tabela 1 apresenta uma classificação dos estudos selecionados quanto a essas abordagens, as quais são explanadas a seguir.

Tabela 1. Classificação dos estudos selecionados quanto às abordagens para integração de dados heterogêneos em cidades inteligentes.

Abordagem	Estudos
Consolidação de dados	E2, E3, E5, E6, E7, E8, E9, E11, E12, E14, E15, E16, E17, E19, E20, E21, E22, E23, E24, E25, E26, E27
Federação de dados	E1, E7, E28
Propagação de dados	E1, E7, E10, E13, E18, E20, E26
Vocabulário controlado	E1, E2, E3, E4, E5, E6, E7, E8, E9, E10, E11, E13, E15, E17

A **consolidação de dados** consiste coletar dados de diferentes fontes e armazenalos em uma única base de dados e, feito isso, tais dados podem disponibilizados às aplicações e usuários. Essa abordagem tem como principal vantagem a sua simplicidade, dado que as aplicações não precisam lidar diretamente com as múltiplas fontes de dados. Além disso, uma vez que os dados estejam localizados em um único local, eles podem ser mais facilmente analisados pelas principais ferramentas e *frameworks* de análise de dados disponíveis atualmente. Como desvantagem, nessa abordagem existe uma espécie de atraso entre a produção da informação e a sua disponibilização e, dependendo da fonte de dados, da natureza destes e da forma como esses dados são consolidados, esse atraso pode variar de alguns segundos a até alguns dias. Os estudos E14 e E16 implementam essa abordagem por meio de *data extractors*, componentes conectados às diversas fontes de dados presentes na cidade e responsáveis por extrair dados de portais de dados abertos, redes sociais, sistemas de monitoramento de transito, etc. Uma vez extraídos, os dados são armazenados em bases de dados centralizadas e, em seguida, consolidados para posterior análise e disponibilização às aplicações.

A abordagem de **federação de dados** envolve uma camada de abstração sobre as diversas fontes de dados que fazem parte da federação, camada essa que oferece uma visão única e organizada acerca dos dados armazenados na camada subjacente. Quando uma aplicação requisita algum dado, a camada de abstração encarrega-se de recuperá-lo a partir da base de dados onde esse dado está armazenado, adaptá-lo ao modelo de dados comum e o retornar para o usuário e/ou aplicação. Esse tipo de abordagem é implementado, por exemplo, pelo estudo E28. O estudo apresenta um *framework* que permite a realização de consultas à informações localizadas em arquivos, sistemas, plataformas e bases de dados heterogêneas. Para isso, o *framework* conta com uma camada de abstração que oferece uma interface única e padronizada por meio da qual os usuários e aplicações podem ter acesso aos dados localizados nos ambientes heterogêneos. Quando um usuário requisita uma informações por meio da interface, o *framework* encarrega-se de localizar a fonte de dados, adquiri-los e convertê-los para o modelo de dados comum utilizado para que então sejam retornados ao usuário.

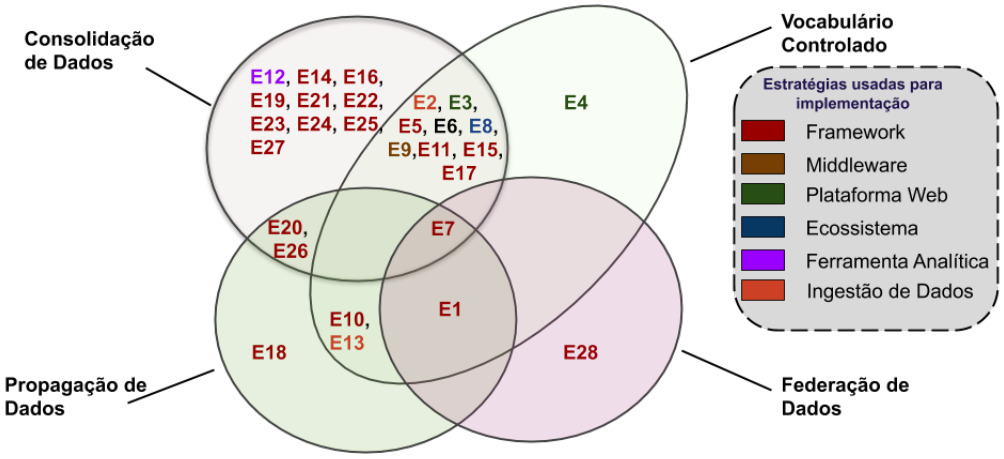

Figura 4. Aboradagens e estratégias para integração de dados heterogêneos em cidades inteligentes.

A abordagem de **propagação de dados**, por sua vez, consiste em mover dados de diversas fontes para bases de dados especificas de forma automática, geralmente com base em eventos. Nessa abordagem, os dados são transferidos para o destino de forma síncrona ou assíncrona. Diferente da estratégia de consolidação de dados, nessa abordagem geralmente o atraso entre a produção e a disponibilização da informação é menor, uma vez que a propagação dos dados é feita de forma automática e disparada por gatilhos/eventos. Esse tipo de abordagem é adotado, por exemplo, pelo estudo E20. O estudo apresenta um *framework* que oferece suporte a integração, análise e disponibilização em tempo real de dados heterogêneos. Para isso, ele conta com um componente responsável por extrair informações de diversos sistemas e dispositivos presentes na cidade.

Por fim, a abordagem por **vocabulário controlado** consiste no uso de convenções ou padrões para organização dos dados trocados entre diferentes sistemas, de modo que todos adotem esse vocabulário comum. Essa padronização é geralmente feita por meio de modelos de dados e/ou ontologias bem definidos. Esse tipo de abordagem é adotado, por exemplo, no estudo E15. O estudo apresenta um modelo de dados implementado por meio de ontologias com o objetivo de descrever semanticamente elementos e indicadores relacionados ao gerenciamento de tráfego em cidades inteligentes. Todo o vocabulário pode ser consumido e reutilizado por sistemas responsáveis pelo gerenciamento do tráfego. Assim, todos os sistemas que adotarem o modelo de dados proposto serão capazes de trocar informações entre si de forma precisa, sem ambiguidades e/ou problemas relacionados à compreensão dos dados.

Implementação da integração de dados. Cada uma das quatro abordagens identificadas nos estudos selecionados pode ser implementada por meio de diversas estratégias. Nos estudos selecionados, elas foram implementadas por meio de (i) *frameworks*, (ii) *plataformas Web*, (iii) *middleware*, (iv) *ecossistemas*, (v) *ferramentas analíticas* e (vi) *ingestão de dados*. A Figura 4 ilustra as abordagens adotadas por cada um dos estudos selecionados, bem como as estratégias utilizadas para sua implementação. É possível observar que o desenvolvimento de *frameworks* foi a estratégia de implementação mais utilizada, correspondendo a aproximadamente 71% (20/28) do total de estudos selecionados. É importante mencionar que alguns dos estudos eventualmente adotaram abordagens mistas para promover a integração de dados.

O desenvolvimento de um ***framework*** consiste em prover um conjunto de componentes, bibliotecas e funções para facilitar o trabalho do desenvolvedor na construção de soluções. Por exemplo, o estudo E19 analisa técnicas e métodos para o desenvolvimento de um *framework* que possibilite a análise, integração e visualização de dados urbanos de diversas fontes heterogêneas, de maneira genérica e reutilizável, permitindo o seu uso em soluções para análise de dados urbanos (transporte e trânsito).

A proposição de **plataformas Web** como forma de promover a extração, integração ou importação de dados produzidos em cidades inteligentes é uma estratégia interessante, uma vez que permite o usuário final manipular informações heterogêneas de maneira transparente, sem se preocupar como é realizada a padronização de tais dados. Essa premissa pode ser observada no estudo E4, que consiste em uma aplicação Web para auxiliar a administração pública da cidade de Catania, na Itália, e seus respectivos cidadãos no ato de relatar e resolver problemas urbanos, tais como buracos em ruas e rodovias, falhas de iluminação pública, etc. Para isso, foi desenvolvida uma interface para os usuários finais (cidadãos) e um sistema para gerenciamento das informações coletadas, as quais são baseadas em dados ligados (*Linked Data*) e adotam princípios da Web Semântica com o propósito de permitir interoperabilidade em níveis sintático e semântico.

A utilização de ***middleware***, no contexto do estudos selecionados, consiste na provisão de uma camada de *software* localizada entre as diversas fontes de dados e as aplicações que as consomem. Tal camada abstrai detalhes relacionados à aquisição e tratamento dos dados e os oferece de forma padronizada às aplicações, além de prover outros serviços adicionais e igualmente importantes, tais como controle de acesso aos dados gerenciados, análise de dados em tempo real, etc. O estudo E9, por exemplo, propõe uma plataforma de *middleware* denominada *Smart Geo Layers* (SGeoL). Para viabilizar a integração de dados heterogêneos em ambientes de cidades inteligentes, o SGeoL adota as abordagens de consolidação de dados e vocabulário controlado. O SGeoL oferece interfaces padronizadas por meio das quais os diversos produtores de informações disponíveis na cidade (dispositivos físicos, sistemas da gestão publica, etc.) podem publicar suas informações e as disponibilizar às aplicações que têm interesse em seus dados. Além dessas funcionalidades, o SGeoL oferece serviços adicionais úteis às aplicações construídas, tais como controle de acesso aos dados gerenciados (autenticação e autorização), suporte ao processamento de dados geográficos e *dashboards* voltados para visualização e análise dos dados gerenciados. Para padronizar a troca de dados, o SGeoL adota o modelo de dados definido no protocolo *Next Generation Services Interfaces* (NGSI) [Bauer et al. 2010].

A abordagem de **ecossistema** tem como ideia principal prover uma arquitetura composta por módulos independentes que, quando reunidos, operam de maneira harmônica. Essa abordagem é utilizada pelo estudo E8, no qual é proposto o *Linked Data Analytics* (LinDA), um ecossistema formado por módulos de: (i) exploração, que permite a coleta de dados de acordo com o *Resource Description Framework* (RDF)[1] e conversão de dados que não estejam nesse formato; (ii) consulta e ligação de dados, fornecendo consultas simples e complexas por meio de um ambiente gráfico, e; (iii) visualização e análise acerca de diferentes categorias de dados, como dados estatísticos, geográficos, temporais, etc. O LinDA foi validado em uma aplicação cujo objetivo era fornecer informações relativas ao efeito dos níveis de poluição atmosférica na vida cotidiana dos cidadãos.

[1] https://www.w3.org/RDF/

O estudo E12 faz uso de **ferramentas analíticas** para análise de dados. Três fases são consideradas nessa estratégia: (i) preparação, em que os dados são coletados e padronizados; (ii) análise, em que os dados são extraídos, transformados e carregados, sendo uma etapa crucial para visualização e tomada de decisão por parte do usuário final, e; (iii) visualização, que proporciona a interação do usuário final com os dados analisados. Como aplicação dessa estratégia, o estudo E12 fez uso de *logs* relativos a dados de coleta de lixo em estações de tratamento e informações sobre o curso de carros de coleta de lixo na cidade de Fujisawa, no Japão.

Por fim, a estratégia de **ingestão de dados**, observada nos estudos E2 e E13, baseia-se na coleta com ou sem armazenamento de informações de diversas fontes, possibilitando a correlação entre elas e as disponibilizando para consumo por outras aplicações. A título de ilustração, no estudo E13, essa ingestão de dados é feita por meio de um aplicativo que opera em segundo plano no dispositivo móvel do usuário com o objetivo de coletar informações acerca de localização e atividades realizadas pelo indivíduo, permitindo assim determinar se ele possui tendência ou apresenta sintomas de depressão. No estudo E2, a coleta de dados sobre transporte urbano resulta em um mapeamento em conformidade com uma ontologia referente a mobilidade em cidades inteligentes, além da persistência em uma base de dados RDF para posterior consulta semântica, visando ao final, fornecer novos serviços e informações aos usuários.

Principais resultados (QP1). As principais abordagens utilizadas para viabilizar a integração de dados em ambientes de cidades inteligentes são consolidação de dados e vocabulário controlado, sendo geralmente utilizadas em conjunto por combinarem caracterísiticas como simplicidade, eficiência e padronização dos dados, além de evitar a necessidade de usuários e aplicações lidarem com fontes de dados e protocolos heterogêneos. Na maioria dos casos, essas abordagens são implementadas por meio de *frameworks* com o intuito de facilitar o desenvolvimento de aplicações.

3.3. Facilidades para Integração, Extração e Importação de Dados

Como anteriormente discutido, as principais abordagens utilizadas para integração de dados nos estudos selecionados consistem em extrair dados a partir de fontes heterogêneas e os centralizar em um armazenamento comum, de modo a serem mais facilmente consumidos pelas aplicações. Depois de centralizados em um local específico, esses dados são em geral oferecidos às aplicações por meio de interfaces e/ou APIs de alto nível. Com isso, não existe, por parte dos usuários e aplicações, a necessidade de lidar com múltiplas fontes de dados distribuídas e com protocolos e modelos de dados distintos a fim de recuperar os dados requeridos, facilitando assim o desenvolvimento de aplicações e a utilização dos dados disponíveis. A partir da análise dos estudos selecionados, as facilidades identificadas para viabilizar a integração de dados heterogêneas em ambientes de cidades inteligentes se constituíam basicamente de (i) APIs/interfaces padronizadas para consulta e importação de dados e (ii) utilização ou criação de modelos de dados.

No que diz respeito à adoção de modelos de dados, três quartos do total de estudos selecionados (75%) fazem uso de modelos de dados para padronizar a troca de informações, como mostra a Tabela 2. Nesses casos, o uso de modelos de dados padroniza a troca de informações, contribuindo para a reutilização das mesmas por diferentes sistemas. Observou-se que em nove estudos os autores definiram modelos de dados próprios para representar as informações coletadas, o que acaba de certa forma dificultando o reuso

Tabela 2. Classificação dos estudos selecionados quanto a modelos de dados adotados.

Modelo de dados adotado	Estudos
Próprio	E1, E2, E3, E4, E5, E6, E7, E10, E11, E12, E15, E17, E22, E28
Consolidado	E8, E9, E13, E16, E24, E19

dessas abordagens, uma vez que os dados não estão representados em modelos de dados já consolidados na indústria ou na literatura. Em apenas seis estudos os autores utilizaram modelos de dados já consolidados para representar as informações coletadas, a exemplo de NGSI (utilizado pelo estudo E9), MAnto [de Marina et al. 2018] (utilizado pelo estudo E24) e *General Transit Feed Specification* [Wong 2013] (utilizado pelo estudo E19).

Em relação aos modelos de dados utilizados, foi possível também observar que treze estudos propuseram ou adotaram modelos de dados que proveem suporte semântico para descrição e consulta de informações. Nesses estudos, ontologias como *Semantic Sensor Network*[2] (adotada pelo estudo E22), MAnto (adotada pelo estudo E24) e *CityGML*[3] (adotada pelo estudo E28), além da linguagem OWL[4] (adotada pelos estudos E15 e E17) e modelos de dados baseados em RDF (adotados pelos estudos E2, E10 e E11), são utilizados no intuito de prover interoperabilidade semântica para os dados adquiridos, de forma a facilitar o uso e entendimento destes pelos diversos sistemas da cidade.

Principais resultados (QP2). As principais facilidades oferecidas para viabilizar a integração de dados em ambientes heterogêneos são a provisão de interfaces padronizadas e o uso de modelos de dados padronizados. O uso de interfaces padronizadas evita que os desenvolvedores de aplicações tenham que lidar com fontes de dados distintas, diversos protocolos e representações de dados heterogêneas. Essa abordagem tem sido utilizada juntamente com a adoção de modelos de dados padronizados para representação dos dados.

4. Desafios de Pesquisa e Desenvolvimento

Integrar dados heterogêneos de forma a possibilitar o uso destes de forma facilitada em ambientes de cidades inteligentes não é uma tarefa trivial. Esta seção apresenta uma lista não-exaustiva dos principais desafios apontados pelos estudos selecionados com relação à extração, integração e importação de dados heterogêneos, bem como oportunidades de pesquisas que visem a facilitar essa integração.

Relacionamentos entre os dados. A partir da análise dos estudos selecionados, foi possível identificar lacunas no que se refere a estratégias e soluções para o estabelecimento de relacionamentos entre dados heterogêneos de diferentes fontes. As estratégias e soluções apresentadas nos estudos E2, E7, E10 e E11 fazem uso principalmente de ontologias e tecnologias da Web Semântica no intuito de permitir análises e consultas nas mais diversas fontes e contextos de dados em cidades inteligentes. Uma direção potencial para endereçar essas questões pode ser a incorporação de conceitos de *Linked Data* [Heath and Bizer 2011], que têm se mostrado efetivos na integração de informações da partir de fontes distintas, bem como na identificação e estabelecimento de relacionamentos entre os dados.

[2]https://www.w3.org/2005/Incubator/ssn/ssnx/ssn
[3]https://www.ogc.org/standards/citygml
[4]https://www.w3.org/OWL/

Heterogeneidade dos dados. De acordo com os estudos selecionados, um dos principais desafios relacionados à integração de dados está na sua natureza heterogênea. Em geral, os dados em ambientes de cidades inteligentes são representados em diversos formatos e são estruturados de diversas formas. Embora todos os estudos selecionados tenham apresentado propostas para integrar dados heterogêneos, elas ainda não conseguem mitigar completamente os problemas de integração de dados em uma cidade inteligente e boa parte das propostas é focada em domínios específicos, como mobilidade, meio ambiente, etc. Esses desafios tendem a ser mitigados com a adoção de modelos de dados e padrões abertos e consolidados. No entanto, como discutido anteriormente, a maior parte dos estudos selecionados optou utilizar modelos de dados próprios. Além disso, há a imprevisibilidade de como os dados possam se apresentar, seja em termos de consistência de informações ou quanto a formato.

Inconsistência dos dados. Inconsistências podem acontecer devido a imperfeições que afetam a qualidade dos dados coletados para a integração, a exemplo de um mesmo atributo/informação estar presente em dois conjuntos de dados distintos apresentando valores distintos. Outra situação diz respeito a casos de instabilidade dos dados extraídos, como reportado em alguns dos estudos selecionados, devido a determinado conjunto de dados apresentar uma volatilidade com relação ao padrão e/ou consistência em seu formato, além da disponibilidade desses dados, o que pode inviabilizar a realização de extração e análise. Em geral, problemas desse tipo são difíceis de se resolver na fase de integração. Dessa forma, estratégias como a definição ou a criação de modelos de dados, como as anteriormente abordadas, podem auxiliar na resolução desse desafio.

Frequência de atualização dos dados. A frequência de atualização dos dados diz respeito ao quão atuais se apresentam os dados coletados, por meio da taxa de atualização e políticas de coleta de novos dados. Essa premissa ocorre no estudo E26, que realiza a extração de dados oriundos de bases de dados de redes sociais, aplicações para dispositivos móveis e *websites* e, para cada uma dessa fontes de dados, são definidas diferentes políticas (tempo real, periodicidade por minutos ou semanas, sob demanda). Contudo, essa frequência de atualização dos dados se torna um desafio para a integração de dados, pois, dependendo do contexto, pode ser necessário um balanço entre buscar informações em tempo real, o que é geralmente atrelado a um alto custo computacional, e obter essas mesmas informações sob demanda ou de forma periódica, demandando assim estudos acerca de otimizações para tal tarefa.

5. Considerações Finais

Este artigo apresentou os resultados de um mapeamento sistemático realizado com o objetivo de prover um panorama do atual estado da arte com relação à extração, integração e importação de dados heterogêneos no contexto de cidades inteligentes. Um total de 28 estudos encontrados em bases eletrônicas de publicação foi sistematicamente analisado para identificar as estratégias, metodologias e facilidades existentes para apoiar e implementar essas atividades. A análise dos estudos selecionados permitiu observar que:

- a maior parte dos estudos faz uso de estratégias de consolidação de dados e vocabulário semântico, com vistas a simplicidade, eficiência e padronização dos dados;
- há uma tendência na utilização de modelos semânticos em combinação com a utilização de RDF, visando uma melhor padronização e integração de informações de dados heterogêneos, fazendo com que o uso desses tipos de modelos de dados seja

um caminho para desenvolvimento de novas soluções/estratégias para integração de dados heterogêneos em cidades inteligentes, e;

- há um *trade-off* no que diz respeito à frequência na atualização dos dados e como se pode otimizar tal atualização dependendo do contexto de demanda para coleta de dados (tempo real, sob demanda ou periódica) para determinadas aplicações.

O mapeamento sistemático realizado também permitiu identificar alguns desafios de pesquisa e desenvolvimento com relação às etapas de extração, integração e importação de dados heterogêneos. Dentre esses desafios estão questões relacionadas (i) ao relacionamento entre os dados, (ii) à natureza heterogênea e por vezes inconsistente destes e (iii) à frequência de atualização dos dados produzidos pelas mais diversas aplicações e dispositivos que compõem o ecossistema de uma cidade inteligente.

Referências

Bauer, M. et al. (2010). The Context API in the OMA Next Generation Service Interface. In *Proceedings of the 14th International Conference on Intelligence in Next Generation Networks*, USA. IEEE.

d'Aquin, M., Davies, J., Motta, E. (2015). Smart cities' data: Challenges and opportunities for semantic technologies. *IEEE Internet Computing*, 19:66–70.

de Marina, P. C. G., Alonso, A. S., Sánchez, B. V., Barca, J. M. C., Quintero, C. E. C. (2018). Towards smart public transport data: A specific process to generate datasets containing public transport accessibility information. In *Proceedings of the the Third International Conference on Universal Accessibility in the Internet of Things and Smart Environments*, páginas 66–71, USA. IARIA.

Heath, T. Bizer, C. (2011). *Linked Data: Evolving the Web into a global data space*. Morgan & Claypool Publishers, USA.

Kitchenham, B. A. Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. techreport, Keele University/University of Durham, United Kingdom.

Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M. (2008). Systematic mapping studies in Software Engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, páginas 68–77, United Kingdom. British Computer Society.

Silva, L., Lima, J. A., Cacho, N., Adachi, E., Lopes, F., Cavalcante, E. (2018). Integração, relacionamento e representação de dados em cidades inteligentes: Uma revisão de literatura. In *Anais do I Workshop Brasileiro de Cidades Inteligentes*, páginas 27–36, Brasil. SBC.

Souza, A. et al. (2017). A data integration approach for smart cities: The case of Natal. In *Proceedings of the Third IEEE International Smart Cities Conference*, USA. IEEE.

von Landesberger, T., Brodkorb, F., Roskosch, P., Andrienko, N., Andrienko, G., Kerren, A. (2016). MobilityGraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE Transactions on Visualization & Computer Graphics*, 22(1):11–20.

Wong, J. (2013). Leveraging the general transit feed specification for efficient transit analysis. *Transportation Research Record*, 2338(1):11–19.

Zhang, H. Babar, M. A. (2010). On searching relevant studies in Software Engineering. In *Proceedings of the 14th International Conference on Evaluation and Assessment in Software Engineering*, páginas 111–120, USA. ACM.

Apêndice A. Estudos selecionados

[E1] Montanelli, S., Castano, S., Genta, L. (2012). Multi-Web, event-centric urban information integration. In *Proceedings of the 8th International Conference on Signal Image Technology and Internet Based Systems*, páginas 852–859, USA. IEEE.

[E2] Bellini, P., Nesi, P., Rauch, N. (2014). Knowledge base construction process for smart-city services. In *Proceedings of the 19th International Conference on Engineering of Complex Computer Systems*, páginas 186–189, USA. IEEE.

[E3] Psyllidis, A., Bozzon, A., Bocconi, S., Bolivar, C. T. (2015). A platform for urban analytics and semantic data integration in city planning. In Celani, G., Sperling, D. M., Franco, J. M. S., editores, *Computer-Aided Architectural Design Futures*, volume 527 de *Communications in Computer and Information Science*, páginas 21–36. Springer Berlin Heidelberg, Germany.

[E4] Consoli, S., Recupero, D. R., Mongovi, M., Presutti, V. Cataldi, G., Patatu, W. (2015). An urban fault reporting and management platform for smart cities. In *Proceedings of the 24th International Conference on World Wide Web*, páginas 535–540, USA. ACM.

[E5] Tan, Y., Zhang, C., Mao, Y., Qian, G. (2015). Semantic presentation and fusion framework of unstructured data in smart cites. In *Proceedings of the 10th Conference on Industrial Electronics and Applications*, páginas 897–901, USA. IEEE.

[E6] Kozievitch, N. P., Almeida, L. D. A., Silva, R. D., Minetto, R. (2016). An alternative and smarter route planner for wheelchair users: Exploring open data. In *Proceedings of the 5th International Conference on Smart Cities and Green ICT Systems*, USA. IEEE.

[E7] Puiu, D. et al. (2016). CityPulse: Large scale data analytics framework for smart cities. *IEEE Access*, 4:1086–1108.

[E8] Fotopoulou, E. et al. (2016). Linked Data analytics in interdisciplinary studies: The health impact of air pollution in urban areas. *IEEE Access*, 4:149–164.

[E9] Souza, A. et al. (2017). A data integration approach for smart cities: The case of Natal. In *Proceedings of the 2017 International Smart Cities Conference*, USA. IEEE.

[E10] Yang, Z., Gupta, K., Gupta, A., Jain, R. K. (2017). A data integration framework for urban systems analysis based on geo-relationship learning. In *Proceedings of the ASCE International Workshop on Computing in Civil Engineering 2017*, USA. ASCE.

[E11] Beseiso, M., Al-Alwani, A., Altameem, A. (2017). An interoperable data framework to manipulate the smart city data using semantic technologies. *International Journal of Advanced Computer Science and Applications*, 8(8):68–72.

[E12] Komamizu, T., Nakazawa, J., Amagasa, T., Kitagawa, H., Tokuda, H. (2017). Analytical toolbox for smart city applications: Garbage collection log use case. In *Proceedings of the 2017 IEEE International Conference on Big Data*, páginas 4105-4110, USA. IEEE.

[E13] Yue, C. et al. (2017). Fusing location data for depression prediction. In *Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, USA. IEEE.

[E14] Balasubramani, B. S., Belingheri, O., Boria, E. S., Cruz, I. F., Derrible, S., Siciliano, M. D. (2017). GUIDES: Geospatial urban infrastructure data engineering solutions. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, páginas 1–4, USA. ACM.

[E15] Mejia, D., Villanueva-Rosales, N., Torres, E., Cheu, R. L. (2017). Integrating heterogeneous freight performance data for smart mobility. In *Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, USA. IEEE.

[E16] Rodrigues, D., Boukerche, A., Silva, T. H., Loureiro, A., Villas, L. A. (2017). SMAFramework: Urban data integration framework for mobility analysis in smart cities. In *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*, páginas 227–236, USA. ACM.

[E17] Shivaprabhu, V. R., Balasubramani, B. S., Cruz, I. F. (2017). Ontology-based instance matching for geospatial urban data integration. In *Proceedings of the 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, páginas 1–8, USA. ACM.

[E18] Komamizu, T., Amagasa, T., Shaikh, S. A., Shiokawa, H., Kitagawa, H. (2017). SOLA: Stream OLAP-based analytical framework for roadway maintenance. In *Proceedings of the 9th International Conference on Management of Digital EcoSystems*, páginas 35–41, USA. ACM.

[E19] Fortini, P. M., Davis, C. A. (2018). Analysis, integration and visualization of urban data from multiple heterogeneous sources. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Advances on Resilient and Intelligent Cities*, páginas 17-–26, USA. ACM.

[E20] Sarabia-Jacome, D., Belsa, A., Palau, C. E., Esteve, M. (2018). Exploiting IoT data and smart city services for chronic obstructive pulmonary diseases risk factors monitoring. In *Proceedings of the 2018 IEEE International Conference on Cloud Engineering*, páginas 351–356, USA. IEEE.

[E21] Lu, Y., Misra, A., Sun, W., Wu, H. (2018). Smartphone sensing meets transport data: A collaborative framework for transportation service analytics. *IEEE Transactions on Mobile Computing*, 17(4):945–960.

[E22] Badidi, E., Maheswaran, M. (2018). Towards a platform for urban data management, integration and processing. In *Proceedings of the 3rd International Conference on Internet of Things, Big Data and Security*, páginas 299–306, Portugal. SciTePress.

[E23] Huang, X., Wang, L., Yan, J., Deng, Z., Wang, S., Ma, Y. (2018). Towards building a distributed data management architecture to integrate multi-sources remote sensing big data. In *Proceedings of the 20th IEEE International Conference on High Performance Computing and Communications; 16th IEEE International Conference on Smart City; 4th IEEE International Conference on Data Science and Systems*, páginas 83–90, USA. IEEE.

[E24] Vela, B., Cavero, J. M., Cáceres, P., Cuesta, C. E. (2019). A semi-automatic data–scraping method for the public transport domain. *IEEE Access*, 7:105627–105637.

[E25] Lee, R., Park, M., HwanLee, S. (2019). An advanced IoT data collection service for data-centric smart cities. *International Journal of Innovative Technology and Exploring Engineering*, 8(8S2):323–327.

[E26] You, L., Tunçer, B., Xing, H. (2019). Harnessing multi-source data about public sentiments and activities for informed design. *IEEE Transactions on Knowledge and Data Engineering*, 31(2):343–356.

[E27] Mehmood, H. et al. (2019). Implementing Big Data Lake for heterogeneous data sources. In *Proceedings of the 35th IEEE International Conference on Data Engineering Workshops*, páginas 37–44, USA. IEEE.

[E28] Chaturvedi, K., Kolbe, T. H. (2019). Towards establishing cross-platform interoperability for sensors in smart cities. *Sensors*, 19(3).