# Publicação de Dados Abertos Conectados Sobre os Transplantes Realizados no IMIP

**Aluna: Rayelle Ingrid Vera Cruz Silva Muniz**

E-mail: rivcs@cin.ufpe.br

**Orientadora: Bernadette Farias Lóscio**

E-mail: bfl@cin.ufpe.br

**Universidade Federal de Pernambuco - UFPE**

**Programa de Pós-Graduação em Ciência da Computação – Centro de Informática**

**Nível**: Mestrado

**Mês e ano de ingresso:** março/2017

**Mês e ano previstos para defesa:** março/2019

**Etapas Concluídas:** Créditos em disciplinas, Definição do Problema, Especificação e Referencial Bibliográfico Inicial

**Etapas Futuras:** Finalização da Especificação e Implementação, Realização de Experimentos e Escrita da Dissertação

***Abstract.*** *Universities, governments, companies, startups and many other organizations are increasingly using the Web as the primary means of sharing and generating content. One of the areas that have shown interest in publishing data in open format is the Health's area. With the dissemination of the Semantic Web and the Linked Data principles we can find several clinical studies, relevant analyzes of hospital data and various other valuable information being published in open format. Given this scenario, this work aims the creation of a linked open dataset regarding the transplants performed in IMIP.*

***Resumo.*** *Universidades, governos, corporações, startups e diversas outras organizações estão cada vez mais utilizando a Web como principal meio de compartilhamento e geração de conteúdo. Com isso, uma das áreas que têm demonstrado interesse na publicação de dados abertos é a área da Saúde. Com a disseminação da Web Semântica e dos princípios de Linked Data podemos encontrar diversos estudos clínicos, análises relevantes de dados hospitalares e diversas outras informações de valor sendo publicadas em formato aberto. Diante desse cenário, este trabalho tem por objetivo a criação de um conjunto de dados abertos e conectado a respeito dos transplantes realizados no IMIP.*

***Palavras-Chave:*** *Dados Abertos, Dados Abertos Conectados, Linked Data, Publicação e Consumo de Dados, Dados de transplantes, Dados na Web.*

## 1. Introdução e Motivação

A Web é hoje o principal meio de compartilhamento de informações. O crescimento do número de fontes de dados na Web tem aumentado o interesse de organizações, instituições de ensino e governos a publicar seus dados em formato aberto, além de motivar o desenvolvimento de aplicações e ferramentas que consomem o conteúdo disponibilizado por estas fontes.

O interesse na utilização da Web como plataforma para publicação de dados não é algo novo [Berners-Lee *et al.* 1999, Abiteboul *et al.* 2000]. Porém, nos últimos anos este interesse tem aumentado devido à flexibilidade que a Web provê para a publicação e consumo de dados. Em sites como o DataHub[1] é possível encontrar diversas fontes de dados em formato aberto, o que confirma o grande potencial que a Web tem para a publicação e o consumo de dados e o constante interesse dos provedores de dados em utilizá-la.

A publicação de dados em formato aberto tem gerado benefícios em diversas áreas, como a transparência dos órgãos governamentais, que provê melhor acesso aos dados, participação e colaboração da sociedade em seus governos [W3C Escritório Brasil 2011]. A abertura dos dados também pode contribuir para o avanço da ciência, bem como a geração de novos empregos; pode aumentar, também, a qualidade nos serviços prestados por diversas organizações, além de muitos outros benefícios que podem ser encontrados em [Pires 2015].

Além do setor governamental, uma das áreas que têm demonstrado interesse na publicação de dados abertos é a área da Saúde. Muitas instituições de saúde e/ou provedores de dados independentes, têm criado *datasets* que possibilitam a geração de análises relevantes. Como exemplo, podemos encontrar no DataHub o *dataset* "*Pharmaceutical Drug Spending by Countries*[2]" que possui indicadores sobre o total de gastos com drogas farmacêuticas para saúde por país. Ainda podemos encontrar o *dataset* "*Breast Cancer*[3]" com dados sobre as ocorrências de câncer de mama. Por fim, podemos mencionar o *dataset* "*Cervical Cancer*[4]" com dados sobre as ocorrências de câncer cervical nas mulheres.

Porém, infelizmente em alguns casos os dados a serem publicados em formato aberto encontram-se em sistemas proprietários ou até mesmo em registros físicos. Nesses casos, o processo de publicação requer algumas etapas adicionais que podem envolver desde a digitalização dos dados até a modelagem dos conjuntos de dados a serem publicados.

Nesse contexto, este trabalho tem como objetivo a criação de um conjunto de dados sobre os transplantes realizados no Instituto de Medicina Prof. Fernando Figueira (IMIP)[5]. Atualmente, o cadastro dos pacientes candidatos a um transplante, bem como

[1] https://datahub.io

[2] https://datahub.io/core/pharmaceutical-drug-spending

[3] https://datahub.io/machine-learning/breast-cancer

[4] https://datahub.io/machine-learning/cervical-cancer

[5] http://www1.imip.org.br/imip/home/index.html

os dados dos pacientes transplantados são realizados manualmente. O setor de transplantes do IMIP conta com os dados de mais de 1600 pacientes transplantados e mais de 500 pacientes aguardando na fila por um transplante. Considerando que os primeiros registros foram feitos há mais de 15 anos atrás, o volume de dados disponível é grande e pode ser bastante útil para a realização de análises, como a identificação do perfil dos pacientes transplantados com maior taxa de sobrevivência.

É importante observar que o conjunto de dados, resultado deste trabalho, será criado de acordo com as boas práticas para dados na Web (*Data on the Web Best Practices).* [Lóscio *et al.* 2017], propõe um conjunto de boas práticas com o intuito de produzir conjuntos de dados de qualidade facilitando uma melhor comunicação entre provedores e consumidores de dados. Além disso, será criado um vocabulário para a descrição dos metadados estruturais. O uso desse vocabulário também facilitará a publicação dos dados de acordo com os princípios de *Linked Data*.

O restante deste artigo está organizado como se segue: a Seção 2 introduz alguns conceitos; a Seção 3 descreve a solução proposta; a Seção 4 apresenta a metodologia utilizada para a realização deste trabalho; a Seção 5 discute alguns trabalhos relacionados, e a Seção 6 traz algumas considerações, indicando os próximos passos para sua conclusão.

## 2. Fundamentação Teórica

O conceito de Dados Abertos aplica-se a todo dado publicado na Web disponível para que qualquer usuário possa utilizar, reutilizar e redistribuir esse dado sem qualquer restrição de patentes, propriedade intelectual ou outro mecanismo de controle, estando sujeito, no máximo, a atribuição de autoria.

*Linked Data* (Dados Conectados), por outro lado, é um conjunto de princípios para publicação de dados estruturados. Um dos principais objetivos do *Linked Data* é prover uma Web onde os dados possam estar diretamente ligados com outros dados por meio de *links* RDF, possibilitando a navegação entre diferentes conjuntos de dados e permitindo a realização de inferências.

A integração desses dois conceitos resulta nos dados abertos conectados ou Linked Open Data [Isotani and Bittencourt 2015], tornando mais fácil a manipulação e reutilização dos dados, agregando valor e possibilitando a descoberta de novos dados vinculados. É importante ressaltar que dados conectados não necessariamente precisam ser abertos.

O processo de publicação e consumo de dados na Web envolve várias fases que vão desde a preparação dos conjuntos de dados a serem publicados até o *feedback* sobre os dados utilizados e o refinamento dos dados gerados. Esse conjunto de fases que compõe o processo de publicação e consumo dos dados é chamado de Ciclo de Vida dos Dados na Web.

As fases do ciclo de vida representado na Figura abaixo são brevemente descritas a seguir [Lóscio *et al.* 2015]:


**Ciclo de Vida dos Dados na Web**

1. **Preparação:** A primeira fase começa desde o momento em que há a intenção de se publicar os dados e se estende até a seleção dos dados que serão publicados.
2. **Criação:** Esta etapa é a de extração dos dados de fontes de dados já existentes até a sua transformação para o formato adequado para publicação na Web.
3. **Avaliação:** Esta etapa requer a avaliação de especialistas da área em que se quer publicar os dados, a fim de que eles possam certificar a qualidade dos mesmos.
4. **Publicação:** Após a avaliação dos especialistas, os dados serão disponibilizados de forma pública na Web, sendo importante a garantia ao usuário de que os dados serão atualizados de acordo com uma frequência pré-determinada, a qual deverá ser disponibilizada juntamente com os dados.
5. **Consumo:** Nesta fase os dados estão disponíveis para serem utilizados para a criação de visualizações, como gráficos, bem como aplicações que permitam a realização de análises sobre os dados.
6. ***Feedback:*** Uma das fases e maior importância, pois é a partir do *feedback* dos usuários que é possível identificar melhorias e realizar correções nos dados previamente publicados.
7. **Refinamento:** Esta fase compreende todas as atividades relacionadas a adições ou atualizações nos dados que já foram publicados. É de suma importância garantir a manutenção e correção dos dados, de acordo com os *feedbacks* recebidos pelos consumidores, a fim de oferecer maior segurança para os consumidores dos dados.

## 3. Solução Proposta

O objetivo geral deste trabalho é a publicação de um conjunto de dados abertos conectados sobre os transplantes realizados no IMIP, a ser criado de acordo com as melhores práticas propostas em [Lóscio *et al.* 2017], bem como a criação de um vocabulário para a descrição dos metadados estruturais.

Atualmente, as informações sobre os pacientes transplantados e os que estão na fila de espera para o procedimento de transplante estão distribuídas em vários arquivos físicos que, no caso das fichas de pacientes transplantados, possuem dados gerais de cada paciente, de seus doadores, informações importantes para o transplante, o cirurgião que realizou o procedimento, além de informações sobre cada dia de pós-operatório desses pacientes. Já nas fichas dos pacientes que aguardam um transplante possui dados

gerais desses pacientes, dados imunológicos e uma série de exames que avaliarão a taxa de rejeição ou sobrevivência de um órgão no paciente.

Levando em consideração que atualmente o IMIP possui registros de mais de 1600 pacientes transplantados e mais de 500 pacientes aguardando na fila por um transplante, a realização de análises sobre esses dados torna-se muito custosa e complexa para um gestor realizar. A criação de um conjunto de dados abertos com esses registros beneficiará todos os envolvidos nas etapas do processo de realização de um transplante, pois auxiliará na visualização, geração de análises sobre esses dados, e realização de projetos de pesquisas.

As etapas de criação do *dataset* seguirão as etapas do ciclo de vida dos dados na Web, como descrito a seguir. A etapa de **Preparação** abrangerá um estudo e análise dos dados disponíveis sobre os pacientes de pré-transplante e transplantados. Nessa fase será necessário a digitalização dos documentos disponíveis e uma análise dos dados que são relevantes e dos dados que podem ser publicados em formato aberto, respeitando a privacidade de cada paciente, visto que muitos dados presentes nas fichas são sigilosos. A etapa de **Criação** diz respeito à modelagem dos conjuntos de dados e à criação do vocabulário para descrição dos dados. Nessa etapa utilizaremos uma ferramenta ou *api* para auxiliar na geração das triplas *RDF*, seguindo os princípios de *Linked Data.* Desta forma o dado estará aberto e conectado, possibilitando a realização de inferências e a criação de *links* com outros *datasets*.

Na etapa de **Avaliação** os especialistas do IMIP irão avaliar amostras dos dados para certificar a qualidade dos mesmos e de que nenhum dado infrinja a privacidade de cada paciente. Na etapa de **Publicação,** o conjunto de dados será disponibilizado em um portal de dados abertos que fará uso de um SGDW para o gerenciamento dos conjuntos de dados proposto em [Oliveira *et al.* 2018]. Esse SGDW é uma solução mais completa, pois permite a definição, criação, manutenção, manipulação e compartilhamento de conjuntos de dados na Web, enquanto que as soluções atualmente disponíveis se concentram mais na catalogação de conjuntos de dados. Publicaremos o *dataset* nos diversos níveis de abertura dos dados, seguindo os princípios de *Linked Data* e as melhores práticas para publicação de dados na Web.

Na etapa de **Consumo,** os dados estarão disponíveis em um portal para o consumo desses dados, a fim de que eles possam ser utilizados para a criação de visualizações, como gráficos, estatísticas, bem como para a realização de análises sobre os dados, como a quantidade de pacientes transplantados que precisaram ser readmitidos em um certo período de tempo após o procedimento de transplante. A etapa de **Feedback** é uma das etapas de suma importância no ciclo de vida dos dados, em especial para este trabalho, pois é nela que receberemos *feedback* sobre o conjunto de dados gerado de forma a mantê-lo em constante atualização e melhoria, bem como a escolha de novos dados a serem publicados. A última etapa é a de **Refinamento** e reflete o momento para correção de possíveis erros e, se for o caso, repetir todo o processo do ciclo de vida, desde a avaliação até um novo refinamento dos dados.

## 4. Metodologia

O processo de criação desse trabalho divide-se essencialmente em 4 etapas. A primeira etapa foi dedicada ao levantamento do estado da arte, onde encontramos alguns

conjuntos de dados relacionados à área de saúde e que estão em formato aberto e/ou conectado. A segunda etapa diz respeito ao estudo das boas práticas para publicação de dados, a fim de aplicá-las ao nosso projeto. A partir daí será possível a criação e publicação do *dataset* de acordo com essas boas práticas.

A terceira etapa desse trabalho é onde avaliaremos o *dataset* criado com o auxílio de especialistas na área. A quarta e última etapa é quando divulgaremos os resultados obtidos por meio da escrita de artigos e pela escrita da dissertação como requisito parcial para obtenção do título de Mestre.

## 5. Trabalhos Relacionados

Apesar do grande número de *datasets* com dados abertos e conectados, ainda são poucos na área de saúde, especificamente, em comparação com *datasets* sobre dados governamentais, por exemplo. Além disso, muitos dos *datasets* que contém dados sobre saúde estão em formato aberto, segundo a classificação de abertura dos dados proposto por Tim Berners-Lee, porém nem todos estão em formato conectado.

Além do site DataHub mencionado anteriormente, também podemos citar o site HealthData.gov[6] que incorpora 125 anos de dados de saúde nos EUA, onde podemos encontrar conjuntos de dados fornecidos por agências em todo o Governo Federal, bem como as ferramentas e aplicativos para manipulação e processamento de dados. É importante ressaltar que o DataHub possui conjuntos de dados abertos, porém nem todos estão conectados. Já o HealthData.gov possui conjuntos de dados conectados, porém nem todos são abertos.

Numa busca rápida no HealthData.gov, encontramos 2.737 *datasets* divididos entre os seguintes tópicos: *Health, State, National, Medicare, Hospital, Quality, Community, Inpatient*. Ao restringirmos a busca para *datasets* que possuam dados sobre "Transplant" ou "Transplantation" o resultado retorna apenas 7 conjuntos de dados[7] [8] [9] [10] [11] [12], tendo um deles encerrado. Destes 7 conjuntos de dados, podemos destacar 3 que possuem algumas características ou análises similares às que serão desenvolvidas neste trabalho.

O primeiro é o *dataset* "*Surgical Site Infections (SSIs) for Operative Procedures in Healthcare*[7]" que contém dados sobre infecções em locais cirúrgicos relatadas por um hospital para o Center for Disease Control and Prevention (CDC) e para o National Healthcare Safety Network (NHSN). O segundo *dataset* é o "*Incidence of Lung Transplants in the Medicare Population*[9]" com análises da incidência de transplantes de

[6] https://www.healthdata.gov/

[7] https://www.healthdata.gov/dataset/surgical-site-infections-ssis-operative-procedures-healthcare

[8] https://www.healthdata.gov/dataset/central-line-associated-bloodstream-infections-clabsi-healthcare

[9] https://www.healthdata.gov/dataset/incidence-lung-transplants-medicare-population

[10] https://www.healthdata.gov/dataset/medicare-ffs-30-day-readmission-rate-puf

[11] https://www.healthdata.gov/dataset/hcup-national-nationwide-inpatient-sample-nis-restricted-access-file

[12] https://www.healthdata.gov/dataset/hrsa-data-warehouse

pulmão na população do Medicare[13] nos últimos anos. O terceiro *dataset* é o "*Medicare FFS 30 Day Readmission Rate PUF[10]*" que contém dados da Taxa de Readmissão Hospitalar (PUF). Este dataset permite analisar as causas em que um beneficiário do Medicare é internado em um hospital dentro de 30 dias da data de alta após uma estadia anterior.

Além desses conjuntos de dados mencionados acima, podemos encontrar outros na literatura, porém eles tratam especificamente de mamografias, e/ou de algum tipo de câncer. Os poucos *datasets* que foram encontrados no filtro de "transplante" ou "doação de órgãos" não tratam especificamente do paciente em si, mas das causas possíveis de complicação e afins.

## 6. Considerações Parciais e Trabalhos Futuros

Este trabalho propõe a criação de um dataset aberto e conectado sobre os dados de pacientes dos transplantes realizados no IMIP. O *dataset* a ser disponibilizado caracteriza-se como importante fonte de informação para diversos novos estudos na área de saúde e afins, além de auxiliar na visualização e geração de análises sobre os dados publicados. Atualmente, o projeto encontra-se na fase de preparação do *dataset*. Considerando que os dados a serem publicados estão dispostos em diversos arquivos físicos, será necessário a inserção desses dados em uma base de dados para que então possamos analisar, juntamente com o médico responsável, quais dados podem estar em formato aberto, visto que muitos deles possuem informações particulares de cada paciente.

## Referências

Abiteboul, S., Buneman, P., and Suciu, D. (2000). *Data on the Web: from relations to semistructured data and XML.* Morgan Kaufmann.

Berners-Lee, T., Connolly, D., and Swick, R. R. (1999). *Web architecture: Describing and exchanging data.* Disponível em: <https://www.w3.org/1999/04/WebData>. Acesso em: 27 mai. 2018.

Isotani, S. and Bittencourt, I. I. (2015). *Dados Abertos Conectados*. novatec, 1st edition.

Lóscio, B. F., Burle, C., and Calegari, N. (2017). *Data on the Web Best Practices*. Disponível em: <https://www.w3.org/TR/dwbp/>. Acesso em: 10 mai. 2018.

Lóscio, B. F., Oliveira, M. I. S., and Bittencourt, I. I. (2015). *Publicação e consumo de dados na web: Conceitos e desafios*. In *Minicurso SBBD*.

Oliveira, L. E. R. A., Oliveira, M. I. S., Santos, W. C. R., and Lóscio, B. F. (2018). *Data on the Web Management System: A Reference Model.* In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age* (p. 2). ACM.

Pires, M. T. (2015). *Guia de Dados Abertos*.

W3C Escritório Brasil and Laboratório Brasileiro de Cultura Digital (2011). *Manual dos Dados Abertos: Desenvolvedores.*

[13] https://www.medicare.gov/