Ciência Aberta: o papel dos metadados na descoberta de conhecimento

Open Science: the role of metadata in the knowledge's discovery

Ana Alice Baptista[a]

Transcrição da apresentação

Para mim é bastante diferente, mas é muito bom, é um prazer imenso estar aqui, a primeira coisa que eu queria dizer é agradecer a Universidade Federal da Paraíba, ao evento o convite. Cumprimentar meu colega, os que eu conheço e os que eu não conheço, cumprimentar também a audiência e desejar a todos um excelente Workshop.

Queria dizer também que gostei muito das falas da mesa, gostei muito de saber que os anais do Workshop vão ser disponibilizados assim dessa forma. É uma almofada de ar fresco, porque nós já temos as tecnologias para fazer isso a anos, e a comunidade científica é muito conservadora nos seus processos, não é conservadora em suas investigações mas em seus processos na minha opinião é muita conservadora.

Nós já mudamos e temos as tecnologias suficientes para fazer muitas coisas e não fazermos, e portanto é muito bom e fico muito feliz, dou toda autorização para fazerem a transcrição ou fazerem que quiserem porque é disso mesmo que precisamos, precisamos pôr as coisas a mexer, abalar um pouquinho as estruturas e fazer de forma diferente, porque não é pela dificuldade tecnológica que não fazemos. Essa era uma das coisas que eu ia dizer e portanto certamente estou aqui com essa perspectiva para contribuir para algo de bom, para que as coisas que fazemos estejam disponíveis e possam ser usadas, é disso que eu venho a falar hoje.

Eu vou falar de ciência aberta, da minha perspectiva de ciência aberta, eu tenho trabalhado ultimamente nas temáticas dos dados fundamentais abertos com a consciente aberta, na Europa eu tenho sentido que a ciência aberta está um pouco ao horizonte de algumas coisas, e o nível de tecnologias, de perspectivas de organização da informação, da minha perspectiva os dados fundamentais abertos estão a caminhar na ciência aberta.

Mas a ciência aberta também já está a querer a pegar o comboio. Acho que alguns anos atrás eu ficava muito na questão do acesso infra ao acesso aberto que é muito importante, mas muito na questão do repositório e pouco além dos repositórios. Portanto vou dar minha perspectiva sobre isso, vou falar do Linked Open Data ou dados linkados abertos, acho que no Brasil seriam dados abertos vinculados, metadados e claro para aquilo que me convidaram para vir aqui que é papel dos metadados na ciência aberta e que tem a ver com essas coisas todas.

E vou começar falando o que significa estar aberto. Estar aberto eu vou buscar a definição da Open Knowledge Foundation que diz que aberto significa que qualquer pessoa pode livremente usar, modificar, compartilhar, para qualquer propósito, portanto está aberto para outro acessar.

No Portal Português de Ciência Aberta tem uma frase muito interessante que é essa "o conhecimento é de todos e para todos". Está de acordo também com o que diz a Open

a Universidade do Minho (Uminho). E-mail: analice@dsi.uminho.pt. ORCID: https://orcid.org/0000-0003-3525-0619. Currículo: http://www.degois.pt/visualizador/curriculum.jsp?key=4103065722022437

Knowledge Foundation o conhecimento e para todos e não somente para alguns. E também com o que o senhor vice-diretor vinha a dizer, e o professor Ricardo e professor Guilherme, na questão da distribuição do conhecimento, ele não pode ficar enclausurado, precisa ser distribuído, disseminado por todos.

Ciência aberta, estamos a falar sobre o que significa estar aberto, está aberto não é só estar disponível, é mais do que isso, é poder ser utilizado. Na Ciência Aberta estamos a falar de partilha sem reservas, de informação, isso aqui já é um olhar meu, sobre o que está no portal português de ciência aberta.

Não estamos apenas a falar de partilha de documentos, de partilha de dados, estamos a falar de partilha de processos também do processo científico, na verdade é da cultura da informação dos processos científicos, expandindo um conceito de responsabilidade social.

Portanto, o que aparece no portal da ciência aberta português fala que publicações e dados abertos, a investigação e inovação abertos, eu botei aqui processos de investigação e inovação abertos, redes abertos, ciência e ciência aberta.

Eu acho que aqui nas publicações não são só publicações, portanto eu não compilei nada sobre mas eu acho que aquelas publicações não são apenas textos, temos mais coisa, portanto temos os vídeos, áudios, fotografias, o que tivermos, desde que sejam carácter científico. Eu mudei então em vez de chamar de publicações vamos chamar de fatos, mas de enquanto estiver a falar sobre publicações estamos a enfocar no texto, e nós podemos fazer muito mais.

Desmistificar os artefatos e os processos, portanto chega da gente ter as coisas disponíveis online, vou dar aqui alguns exemplos. Este aqui são vídeos, e screenshots de vídeos do periódico de vídeos, de um artigo que publicado em texto e possui um vídeo acompanhado. Este aqui é também uma captura de tela da informação e dados que estão disponíveis no European Portal de dados de ciência e tecnologia, portanto são dados científicos que estão ali.

A gente ter só disponível o vídeo é suficiente para encontrar esse vídeo, é o suficiente para conseguir manipular esses dados? Nós conseguimos manipular esses dados, e depois de entrara ali não consegui compreender nada daquilo. Os vídeos estão disponíveis, mas precisamos de mais um bocadinho além de publicar os artefatos na Web e achar que estamos a fazer ciência aberta.

Tem fundação do suporte que não são facilmente encontrados não são pesquisáveis. Então por que vamos abrir se eles estão ali mas não são encontrados e nem pesquisáveis, não dá para fazer nada, qual as necessidades de abrir.

Outro exemplo, entrando dentro dos dados fui buscar dados que a gente consegue ler, esse aqui é em espanhol, e me diz uma coisa, para compreender o que significa entrar aqui dentro dessa ciência, entrar e compreender o que são aqueles dados, para nós humanos que sabemos ler francês, espanhol, e conseguimos saber o significado daquilo que está escrito, não é fácil.

Eu tentei e não tenho a certeza de ter conseguido compreender aquilo, e também embora aqui seja um pouquinho mais fácil, eu tenho número de avaliadores na gestão pública mas eu não sei o que é este CIA, portanto há aqui bastante coisa que a gente não sabe o suficiente para conseguir interpretar esses dados de forma fidedigna, por mais que entenda o que está aqui escrita, aqui eu tenho o portal, etc., várias parcelas, mas por mais que a gente entenda aquilo não conseguimos interpretar bem aqueles dados.

E nós somos humanos, nós temos capacidade de interpretação, agora tendo esses dados abertos e máquina virem buscar esses dados e tentar aproximar esses dados automaticamente sem intervenção humana, ou com pouca intervenção humana, com outros dados não consegue fazer.

Portanto muitos dos artefatos que tentamos disponibilizar não são interpretáveis, e facilmente utilizados por causa disso. A minha utilização dos dados está diretamente ligada com coisas automáticas ou mais automáticas possível, está diretamente ligada a interoperabilidade, em particular, à interoperabilidade semântica, que é ser significado, deve ser informação que permita as máquinas interpretar o que está ali. Portanto se essas informações não vem junto com os dados não se consegue interpretar.

Se não se consegue utilizar nem interpretar para que abrimos? Para que disponibilizar? Na verdade precisamos de dados, na ciência aberta lidamos com dados, e dados sobre os dados, e quando falta dados não faltam só dados, faltam também outras coisas. Nós precisamos dos dados e dos catálogos, e tudo com os formatos adequados e significados embutidos, tenho que trazer significado para que eles possam ser interpretados.

Repare aqui, entrevistas foram realizadas com indivíduos que trabalham em instituições de ensino superior. Pelo catálogo eu consigo ter aquelas informações todas sobre os dados, e consigo ter aquela informação conforme processada por máquinas.

Outro exemplo, datasets que conjuntos de dados foram voltados de questionários sobre a necessidade de informação de médicos, portanto os nossos catálogos podem ter coisas mais simples outras mais rebuscadas.

E depois outra coisa que já tem a ver com os dados, entrar nos dados por dentro e ter mais informações, não apenas informações nos catálogos, mas também informação que está nos dados, e se esses dados trouxer significado com ela a gente consegue responder perguntas mais complexas, como por exemplo, que artigos sobre as necessidades de médicos oncologistas apresentam como resultado necessidade de informação sobre diagnóstico?

Aqui estamos a falar dos dados e aqui estamos a falar dos catálogos. Se essa informação vier dos dados e estiver em uma forma processável por máquina e trazer significado a gente consegue fazer perguntas desse gênero, e consegue aproximar isso daqui com dados de outros sítios facilmente.

É preciso poder interpretar facilmente não somente os catálogos mas os próprios dados. Tem havido um esforço grande por tratar os catálogos em Linked Open Data mas os dados também precisam estar em Linked Data, ou múltiplas informações sobre os dados precisam estar nos catálogos, temos que incluir mais informação no catálogo.

Este aqui é o primeiro site que eu pus no início quando tinha feito a definição de ciência aberta, e portanto são estas coisas todas que para mim que trabalho com dados e imagino que para muitas pessoas que estão aqui, nós precisamo a falar de dados, dados e metadados.

O que nós vamos ter de fato são artefatos e dados sobre os artefatos, os dados abertos também são um tipo artefato só que na definição aparecem separados, e dados sobre dados abertos, o processo e dados sobre o processo, redes abertas de ciência e dados sobre redes abertas de ciência.

Ter os processos abertos mas ter os processos abertos significa ter também dados sobre os processos, e estes dados estarem também abertos, e podem ser trabalhados e processados e cruzados com outros dados.

E informação devem ser interpretáveis para humanos e para máquinas, portanto devem ser human-readable e machine-readable. Uma das coisas que eu vejo também é gente que balança entre dois extremos, um é tudo human-readable ou ter tudo machine, nós humanos precisamos também de ler e processar aquela informação e buscar aqueles dados, e se a informação não estiver legível por máquinas também não conseguiremos trabalhar aquilo.

Viemos aqui a uma sigla que vocês já ouviram que é FAIR, que os dados devem ser FAIR, quer dizer Findable, Accessible, Interoperable and Reusable, ou seja, encontrável, acessível, interoperável e reutilizável. Na minha opinião é que tem havido um esforço muito grande aqui no Findable e no Accesible e esforço bastante menor aqui no Interoperable e Reusable, e eu acho que nós estamos na hora de fazer esse esforço. Tem havido algumas iniciativas mas é preciso mais, é preciso um esforço da comunidade para encontrar esses dados que nós temos disponíveis. Eu sei que os dados de catálogo ainda não são interoperáveis, eu sei que nós temos iniciativas mas não chega, vamos por dados linked data nos catálogos.

Quando eu falo que nestas coisas, no FAIR, estou a falar da coleta de metadados, portanto não estamos a fazer as coisas ainda bem-feita. A gente precisa mudar para o paradigma dos dados abertos dos dados em linked data. Por que? O linked data são dados semanticamente interoperáveis através de comunidades de prática, empresas, governo, possuem uma dificuldade de interoperabilidade mas nós temos que caminhar para lá.

Atualmente na minha opinião o cenário que nós temos é esse. Temos artigos, teses, relatórios, que falam muito bem um com os outros, com o protocolo PMH, mas falam mais ou mesmo porque depois vamos ver os metadados e vemos que algumas coisas não cruzam com outras. Depois temos os repositórios científicos com protocolos que também falam uns com outros, e também algumas outras coisas que falam com outras e estão ali.

Então temos ali coisas que falam uns com as outras mas que não falam com quem estão ao lado, atualmente a expressão que tem sido muito utilizada são os silos de dados. E Quando eu falo de interoperabilidade local é para essas coisas falando uma com as outras, é ter um mínimo de interoperabilidade, quando eu vou a Inglaterra eu não falo com os ingleses mas eles vão me reconhecer, e a mesma coisa com os franceses certo, mas entre os franceses e os ingleses eles conseguem resolver a questão. Com o espanhol todos temos essa questão que também estamos conseguindo resolver a questão. A interoperabilidade é isso, é por significado perceptível pelo outro.

Aqui um exemplo da LOD cloud, que significa Linked Open Data cloud, e esses aqui são datasets, são cada vez mais datasets, e a gente consegue ver ali no meio uns especiais que são a DBPedia por exemplo, mas hoje já tem muitos datasets. Esses datasets para estarem aqui tem que ter dados abertos e cumprir as regras do linked data.

O que eu falo é o que o Berners Lee chamou de dados cinco estrelas, em data portals você encontra dados em XLS e CSV, portanto estamos a falar de duas estrelas e tres estrelas. Mas o que isso significa?

Isso significa que o primeiro está na web mas é PDF, que não é facilmente processada. O segundo está na web como dados estruturados, como por exemplo XLS mas é formato proprietário. O terceiro é igual a este mas já não é proprietário, como CSV. O quarto é estar em RDF, que é um formato de base da web semântica. E o quinto é ser linked data, que significa linked data? Linked data significa dados com contexto, dados com significado, ou seja, eu preciso ter os meus dados e preciso não só de ser ligados entre eles, mas ser ligados a outros dados que já existem para darem contexto aos meus dados. E esses dados que já existem eles próprios estão ligados entre eles e estão ligados com outros, e que nós temos esse tal grafo que vocês viram anteriormente aqui destas ligações todas entre dados e datasets.

O que nós falávamos no início da ciência aberta? Dos artefatos e dados, no portal da ciência aberta portuguesa, nós criamos essas coisas todas não aqui mas aqui em cima, no linked data. Por isso precisamos de identificadores para as coisas, em Web Semântica temos que identificar coisas, e os identificadores têm que ser únicos, e depois ligar esses dados todos, para trazer contexto.

Aqui está um exemplo que eu trouxe, reparem aqui, isto aqui é uma parte de um vocabulário feito pelos meus alunos, no âmbito desse trabalho eles tiveram que cuidar de um vocabulário controlado de áreas científicas e importar aquilo em uma tecnologia que se chama SKOS. Então eles fizeram isso e aqui é uma parte, temos aqui um conceito que é um conceito da área de economia que faz parte desse conceito maior que é o 110, que é indivíduos, instituições e metadados.

Este aqui é chamado na descrição de dados com o link. Em vez de pôr economia, que está human-readable mas está aqui machine-readable o link para conseguir ser processado. Quando eu falo de link para outras coisas é isso, nós temos nossos dados e que ligam com outras coisas. Quando a gente for utilizar este vocabulário controlado pode utilizar pode interrogar meus dados como se tivesse em uma grande gama de dados e não como se estivessem confinados.

Aqui trago outro exemplo que é também do meu grupo de investigação, reparem tem informação sobre várias coisas do grupo, quem é líder do grupo, nesse caso não puseram nome mas está aqui o link, é um link ORCID da pessoa que está a frente do grupo de investigação. Quando falamos de linked data é isso, ou seja, não aparece nenhum nome mas aparece para as máquinas processarem o link. A mesma coisa aqui, e aqui outra vez a mesma coisa, apontando para um vocabulário controlado para dizer qual área, isso é o linked data, é ter as coisas ligadas, é trazer contexto.

Se eu puser tecnologia TSI que é tecnologia e sistemas de informação nós compreendemos mas para uma máquina qualquer a processar aquilo em um contexto qualquer não compreende o que está ali, precisa de informação adicional, senão vou processar inglês, francês, português, tudo igual.

Isso é um exemplo de um grafo mas é uma coisa muito pequenina, na LOD Cloud, é um grafo RDF. Tendo isto nós conseguimos ter isto que é que queremos, e a pergunta que eu faço é hoje o professor Ricardo Sant'Ana disse que os anais desse workshop vão ser publicados em uma forma completamente distinta e aberta. Quando eu falo aberta não é estar simplesmente na Web, quando eu falo aberta é vídeo, texto, etc, aberto no sentido de mais larga.

O que eu gostaria muito é que isto aconteça e estar a acontecer, no governo eletrônico isso está lá na frente na Europa já aconteceu, na ciência está mais devagar, mas pode ser que passe a frente do governo eletrônico. Quando digo isto eu falo de iniciativas da Europa, não estou a falar de casos fantásticos da ciência que ocorrem nos EUA, mas eu vejo alguma coisa em escala a acontecer.

A minha apresentação via terminar aqui, fiquem com esses desafios para vocês, pensem e levem isso convosco, vão ter mais apresentações aqui hoje que estão relacionadas com essa temática, e portanto pensem nisso e pensem em fazer.

Mais algumas coisas que eu acho que pode ser interessante para vocês da Dublin Core Metadata Initiative tem um canal do Youtube com alguns vídeos, a maior parte são inglês mas têm também em espanhol, e português. Há um projeto que tem um muitos cursos interessantes, não só em texto mas também em vídeo, e que são linkados, colocados no linked data, portanto podem confiar.

E se quiserem eu tenho um Scribd sobre Web Semântica que às vezes ponho coisas interessantes, podem também consultar que pode ser interessante. Depois chamar atenção para essa recomendação, ontem ainda eu vi o professor Guilherme falou sobre três brasileiros, é sobre guia prático para se ter dados na web. Isto aqui ainda não está tal como a gente gostaria mas é um bom conjunto de boas práticas para dados na web.

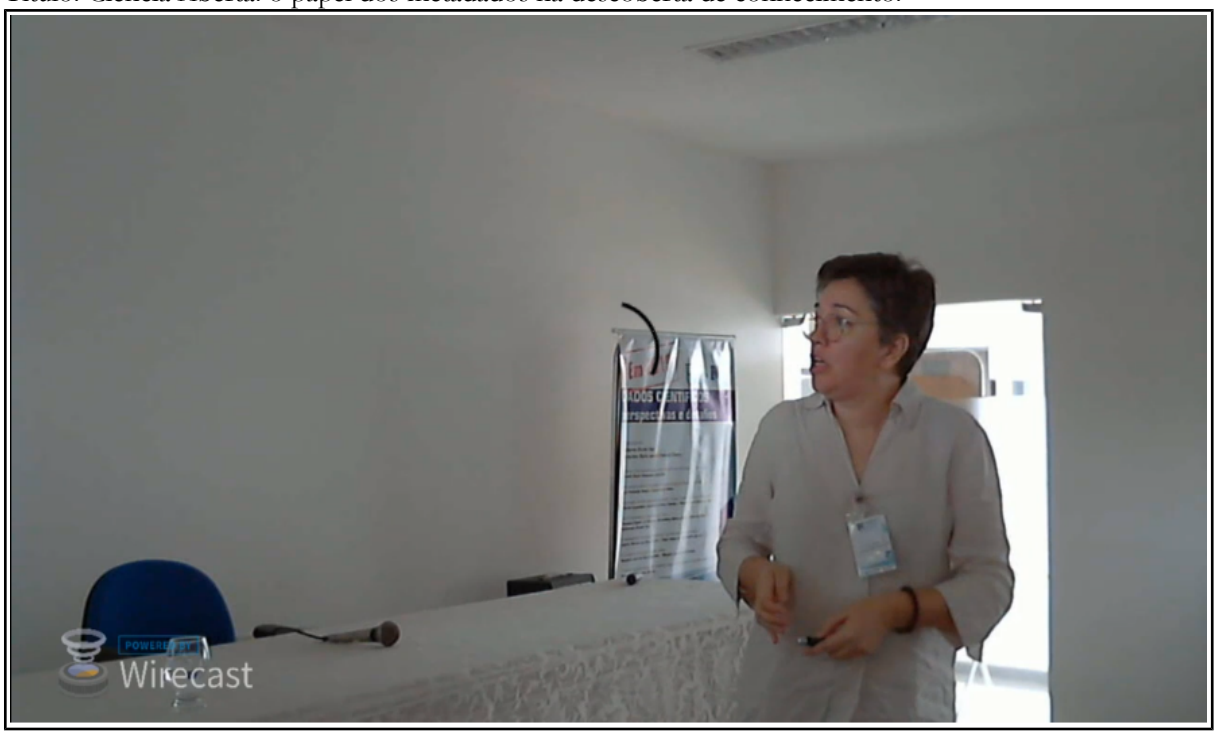
Também chamar a atenção para essa linguagem, que é uma linguagem para fazer espécie de templates para registros RDF, e portanto para fazer uma espécie que funciona como um Schema, que são espécie de formas onde avaliamos nossos dados para ver se nossos dados estão de acordo com as regras definidas, para repositórios, e para quem tem responsabilidade para criar normas pode ser interessante.

Depois por último acompanhar os trabalhos deste Dataset Exchange Working Group que está desenvolvendo uma série de especificações do W3C sobre essas temáticas. Aqui tem várias coisas acontecendo neste grupo começou com um objetivo pequeno e hoje tem quatro ou cinco especificações.

São coisas muito recentes que estão acontecendo e vale a pena que vocês de vez em quando entrem lá para ver o que está a acontecer. E pronto, termino aqui, muito obrigado, e estou aberto a questões.

Vídeo da apresentação

Título: Ciência Aberta: o papel dos metadados na descoberta de conhecimento.



Disponível em: http://dadosabertos.info/enhanced_publications/idt/video.php?id=28