Disponível em www.bad.pt/publicacoes
Paper

Aplicação de análise de conteúdos no desenvolvimento de modelos de metadados para a gestão de dados de investigação

Cristiana Landeira[a], João Aguiar Castro[b], Cristina Ribeiro[c]

[a]*Faculdade de Engenharia da Universidade do Porto, Portugal, crissplandeira7@gmail.com*
[b]*INESC TEC, Faculdade de Engenharia da Universidade do Porto, Portugal, joaoaguiarcastro@gmail.com*
[c]*INESC TEC, Faculdade de Engenharia da Universidade do Porto, Portugal, mcr@fe.up.pt*

Resumo

O aumento da produção de dados, resultantes da investigação científica, desafia a gestão de dados de investigação a desenvolver estruturas de metadados que sustentem a partilha, amplamente defendida pelas agências de financiamento. Nesse sentido, este trabalho consiste no desenvolvimento de modelos de metadados baseados na análise de conteúdos de artigos relacionados com contextos experimentais. De forma a explorar esta abordagem foi desenvolvido um estudo no domínio da Química Sustentável, do qual resultou um conjunto de 60 descritores que foram avaliados por um investigador do domínio. Dos descritores apresentados 53 foram identificados como indo ao encontro das necessidades de descrição do investigador. Através da sessão de avaliação junto do investigador é possível fazer algumas considerações: é notório o aumento do interesse do investigador na colaboração com o curador de dados à medida em que este vê materializado o que é pretendido para a descrição dos dados; verifica-se também a melhoria na comunicação através do conhecimento do domínio adquirido pelo curador. Este processo mostrou a possibilidade de conceber um modelo de metadados num período de tempo reduzido e permitiu a identificação de preferências por parte do investigador relativamente à especificidade ou generalidade dos descritores escolhidos.

Palavras-chave: Gestão de dados de investigação, metadados, análise de conteúdos, química sustentável

Introdução

A expansão de novas ferramentas tecnológicas fomenta a produção de grandes quantidades de dados. Neste contexto a gestão de dados de investigação torna-se cada vez mais um tema de interesse, nomeadamente pelas agências de financiamento que pretendem obter resultados dos projetos que promovem.

Os dados são recursos valiosos resultantes da investigação científica e carecem de investimento para serem geridos de forma a sustentar a partilha, o que é fundamental para rentabilizar os recursos investidos na sua produção. De acordo com a Comissão Europeia (2017 p.5) a partilha influencia a melhoria da qualidade dos dados e favorece a transparência ao permitir a consulta dos resultados por outros investigadores. O incentivo que se dá à investigação colaborativa evita a duplicação de esforços e acelera a inovação. Esta visão é partilhada pelo UK Data Archive (2011 p. 3) que advoga que a partilha dos dados melhora a validação dos métodos de pesquisa e o reconhecimento dos investigadores através da citação.

Apesar dos benefícios, a reutilização dos dados de investigação está dependente do enriquecimento do ponto de vista semântico (Thanos, 2017 p.10). Os dados são difíceis de interpretar quando removidos do contexto que lhes deu origem transformando a partilha de dados num enigma (Borgman, 2012 p. 1). Por isso mesmo, a Comissão Europeia, através das «Guidelines on FAIR Data Management in Horizon 2020», recomenda uma série de princípios que os metadados devem seguir para permitir a descoberta, o acesso, a interoperabilidade e a reutilização dos dados (European Commission, 2016).

Assim, a colaboração entre curadores de dados e investigadores é determinante no desenvolvimento de ferramentas para a criação de metadados. Os primeiros contribuem com a especialização na gestão de informação para garantir as propriedades de preservação e acesso a longo prazo. Os segundos detêm conhecimento específico sobre o domínio em causa (White, 2014 p.40), sem o qual dificilmente se promove a interpretação dos dados. Contudo, para os investigadores o processo de investigação é a atividade prioritária e em muitos casos ainda não estão sensibilizados para a necessidade de gestão e descrição de dados. Por outro lado, também a comunicação entre os curadores e os investigadores tem limitações, sobretudo pela utilização de terminologia diferente. Nem sempre é evidente para os investigadores em que consiste a gestão e dados, ou o que são metadados tornando a interação entre ambos complexa.

A abordagem proposta consiste na seleção e análise de um conjunto de artigos que descrevem experiências similares desenvolvidas num mesmo domínio. Desta forma os curadores de dados recolhem informação para o desenvolvimento de modelos de metadados ajustados às necessidades de cada domínio, sendo requisitado menos esforço dos investigadores no desenvolvimento desses modelos. Um modelo de metadados é composto por um conjunto de descritores que captam valores num determinado contexto.

A análise de conteúdos aplicada a documentos produzidos por investigadores é vista como uma alternativa a uma interação mais extensa com estes, permitindo extrair informação para o desenvolvimento de modelos de metadados (Chao, 2014). Também Wiljes e Cimiano (2012 p.6) defendem a análise de artigos como uma tarefa que o curador deve realizar para que rapidamente adquira conhecimento de um domínio, ainda que o objetivo não seja tornar-se um especialista. Partindo deste princípio, o objetivo é verificar se os curadores de dados podem autonomamente, sem conhecimento prévio nos domínios específicos, compreender e identificar conceitos necessários à descrição de dados nos mais diversos contextos de investigação, sobretudo naqueles onde não estão disponíveis ferramentas para o efeito.

A escolha do domínio da química sustentável para o desenvolvimento do estudo deve-se ao facto da investigação neste domínio dar origem essencialmente a dados do tipo experimental. Estes, quando bem documentados, podem ser reproduzidos se os procedimentos e as variáveis relevantes estiverem documentados (Willis Craig, Greenberg Jane, 2012).

Abordagem metodológica

Este trabalho propõe uma abordagem que consiste na análise pormenorizada de algumas secções de artigos científicos de um domínio escolhido, nomeadamente a *introdução* e secções que relatam a configuração experimental e a metodologia. A escolha de apenas algumas das secções dos trabalhos considerados torna a análise de conteúdos um processo realista. Acresce ainda que da análise integral poderiam resultar valores não necessariamente relacionados com o contexto da produção de dados. Por exemplo, as secções em que são apresentados os resultados não fornecem indicações sobre

o contexto, pelo que se torna desnecessário analisá-las.

A título de exemplo, a metodologia que se propõe é a identificação de «palavras chave» que sirvam como pista para identificar descritores. Assim, quando no artigo o autor se refere à absorbância referindo *«the absorbance at 254 nm»* ou à temperatura a que a amostra foi seca *«the mixture was filtered* and *dried at room temperature»*, os curadores de dados recebem a indicação de que dois dos descritores possíveis são «*Absorbância*» e «*Temperatura de secagem da amostra*».

De modo a validar o trabalho desenvolvido, foi efetuada uma avaliação preliminar com um investigador do domínio da Química Sustentável, especificamente em experiências relacionadas com a remoção de partículas poluentes.

O esforço pedido ao investigador é grande, uma vez que o processo de investigação é prioritário e este tende a não ter disponibilidade para questões relacionadas com a gestão da informação. Nesse sentido, pretende-se com esta abordagem limitar a intervenção do investigador a um envolvimento mínimo possível numa fase inicial, sendo a sua colaboração preponderante na fase de avaliação, onde podem contribuir com sugestões, mostrar preferências, e ainda indicar metadados importantes que não foram previstos, entre outras intervenções.

Para o estudo do domínio da química sustentável começou por se analisar as secções relevantes dos artigos escolhidos (introdução, secções que relatam a configuração experimental e metodologias). A análise de conteúdo é baseada na identificação de palavras relevantes que sirvam de pista para definir os elementos do modelo de metadados. Depois de definida uma proposta de modelo de metadados este é avaliado junto de um investigador do domínio. A Figura 1 representa a abordagem aplicada.

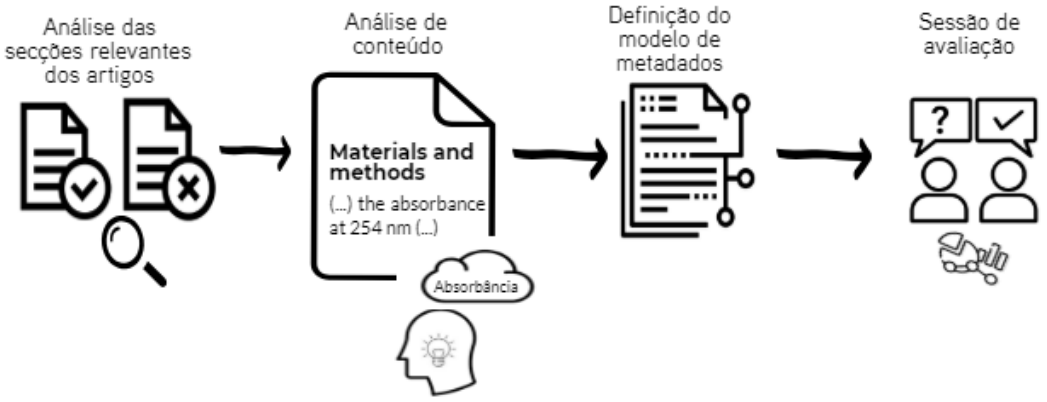

FIGURE 1-ESQUEMATIZAÇÃO DA ABORDAGEM APLICADA

Química sustentável, degradação de partículas poluentes

A atividade humana gera resíduos que se acumulam no meio ambiente, e que influenciam todo o sistema. Os desperdícios resultantes dos processos industriais, da produção agrícola em massa entre tantas outras atividades ao contaminar a água, o solo ou a atmosfera, desencadeiam uma série de problemas para todas as espécies, inclusive a espécie humana. São várias as problemáticas que resultam ou que são amplificadas pela poluição desde as alterações climáticas, escassez de água potável, ou ainda o aparecimento de problemas de saúde. A investigação desenvolvida com o intuito de criar soluções capazes de remover, ou transformar as partículas poluentes é preponderante para tentar reduzir as repercussões causadas pelos agentes poluentes e, portanto, merece ser valorizada.

De forma a tornar a investigação científica neste domínio mais eficiente é necessário que os conjuntos de dados sejam enriquecidos através dos metadados, tornando possível a reprodução das experiências, e a reutilização dos conjuntos de dados, no desenvolvimento de novas investigações.

Assim, para a reconstrução do contexto experimental neste domínio é necessário descrever as propriedades das amostras em estudo, quais os instrumentos selecionados para o desenvolvimento da investigação e sempre que se justificar importa também registar as características e as calibrações do instrumento. Entende-se que qualquer variável que possa ter influência na interpretação dos resultados finais, e que possa auxiliar na descoberta dos dados, é relevante e como tal deve ser descrito.

Metadados que captem valores para os métodos e técnicas utilizados permitem compreender qual o fluxo de trabalho aplicado, possibilitando uma panorâmica da experiência. Ainda, informação relacionada com a duração de determinados eventos, medições e condições ambientais contribui para a qualidade dos metadados.

Avaliação do modelo de metadados

De modo a obter a avaliação do modelo de metadados junto do investigador, foi criado um formulário com o conjunto de descritores identificados. Foi pedido ao investigador para preencher sempre que possível ou fazer um comentário quando não considerasse a inserção de um valor.

Assim, de um conjunto de 60 descritores apresentados 53 foram compreendidos. Dos últimos 38 foram anotados ou aprovados, para os restantes 15 o investigador fez algumas sugestões de melhoria. Por outro lado, dos metadados apresentados, o investigador identifica sete como ambíguos porque: 1) não foram compreendidos; 2) repetem conceitos; 3) não fazem sentido para o investigador. Os resultados podem ser consultados na Tabela 1.

Depois da avaliação efetuada pelo investigador foram discutidas algumas questões. O investigador começou por referir que existe efetivamente uma necessidade de registar informação sobre as experiências e que normalmente esses registos são efetuados em atas, que funcionam como um diário. Nelas são registadas condições de trabalho, condições do desenvolvimento de experiências e ainda acontecimentos inesperados.

Sobre a questão de o modelo de metadados ser ou não suficiente para as necessidades de descrição do domínio, a resposta foi positiva. O modelo capta a informação que habitualmente é registada nas atas. Contudo, afirma que o modelo contém mais elementos dos que são necessários para descrever a experiência.

Quando questionado sobre que descritores não considerados eram relevantes para a descrição do contexto experimental, o investigador sugeriu a necessidade de especificação de algumas concentrações como *Concentração da solução.* Neste descritor seria possível anotar tanto a concentração como a solução em causa não sendo necessário dessa forma criar o descritor *Solução*.

TABELA 1- TABELA DE RESULTADOS DO DOMÍNIO DA QUIMICA SUSTENTÁVEL

Input do curador	**Input do investigador**	
Descritor	**Comentário**	**Proposta**
Composto químico	Causa dúvida, uma vez que tudo é um composto químico assim, não tem especificidade.	Amostra
Coeficiente de transferência de massa volumétrica individual	Deve ser mais específico	Quantidade do composto degradado
Área interfacial	Nem todas as amostras tem uma área interfacial	Distância interfacial

Carência química de oxigénio	Não faz muito sentido desta forma, os dois são parecidos. Assim primeiro é necessário saber qual o gás estudado e depois qual o fluxo da fase gasosa.	Gás em estudo
Taxa de fluxo volumétrico na fase gasosa		Fluxo da fase gasosa
Polifenóis	«Não compreendo, quer saber quantidade ou formula? » Para quantidade deve ser questionado massa/volume do polifenol, se por outro lado, a questão é qual o polifenol usado poderia ser formula do polifenol.	Massa do polifenol
		Formula do polifenol
Potencial de oxidação do agente oxidante	Se o descritor fosse potencial de oxidação a resposta seria a mesma, uma vez que o valor do agente oxidante era captado num dos descritores apresentados.	Potencial de oxidação
Massa Molecular	Importa definir a que corresponde a massa molecular	Massa molecular do polifenol; Massa molecular do gás; (…)
Teor de impurezas	Demasiado específico, normalmente é anotado o grau de pureza e o teor de impureza é obtido através de cálculos.	Grau de pureza
pH do poluente		pH da solução
Tamanho da partícula do poluente		Tamanho da partícula
Tamanho da partícula do catalisador		
Solução aquosa	A palavra «aquosa» repete a ideia de solução, por isso é desnecessário. É importante definir que tipo de solução se trata.	Solução de limpeza; Solução ácida; (…)
Instrumento	Anotaria um nome de um instrumento, contudo a especificação da função do instrumento é mais fácil.	Especificação dos instrumentos como os aprovados.
Método de análise de polifenóis	É sugerido que um descritor mais genérico não restringe tanto.	Método de análise
Descritores compreendidos (aceites)	Elemento químico; Tamanho da cristalite da amostra; Volume do poro da amostra; Carbono inorgânico; Carbono orgânico total; Agente oxidante; Reagentes; Pressão parcial de ozono na fase gasosa; Absorbância; Carbono total; Catalisador; Área da superfície do catalisador; Adsorvente; Área da superfície do adsorvente; Teor de cinzas do adsorvente; Tamanho da partícula do adsorvente; Formula molecular do adsorvente; Amostra de referência; Solução de controlo; Reator de ozonização; Instrumento de medição da absorbância; Instrumento de radiação de luz UV; Instrumento de medição da intensidade da luz; Instrumento de medição do pH; Instrumento de análise dos catalisadores; Instrumento de medição da radiação; eletromagnética; Instrumento para a medição da área superficial; Vaso de reação fotocatalítica; Quantidade da amostra centrifugada; Tempo de ozonização; Tempo de medição da intensidade da luz; Tempo de agitação da suspensão; Temperatura de secagem da amostra; Técnica de remoção de partículas; Técnica de medição da área superficial; Condições atmosféricas; Atividade fotocatalítica; Intensidade da luz solar	
Descritores ambíguos (rejeitados)	Velocidade superficial do gás; Taxa de fluxo de massa de ozono; Concentração interfacial de ozono; Comprimento de onda do catalisador; Absorvência; Instrumento de medição espectral; Quantidade da amostra centrifugada diluída	

Tendo por base a análise de conteúdo efetuada parecia relevante apresentar ao investigador descritores como: *Tamanho da partícula do poluente*, *Tamanho da partícula do catalisador*, *Tamanho da partícula do adsorvente*, ou ainda *Fórmula molecular do adsorvente* e *Método de análise de polifenóis*. A decisão do investigador alternou quando confrontado com estes descritores. Assim, para *Tamanho da partícula do poluente* e *Tamanho da partícula do catalisador* sugeriu apenas *Tamanho da partícula.* Porém para *Tamanho da partícula do adsorvente* e *Fórmula molecular do adsorvente* o investigador anotou respetivamente «80nm» e «carbono ativado», sem fazer qualquer referência. Por sua vez para *Método de análise de polifenóis* sugeriu que o descritor deveria ser mais genérico, apenas Método de análise, de modo a não restringir os valores a registar.

Depois de analisar com atenção o modelo de metadados o investigador sugeriu que o processo do curador pode ser mais eficiente, se este aceder à informação dos instrumentos utilizados nas investigações. Os instrumentos têm campos predefinidos que os investigadores têm de preencher, à semelhança do modelo de metadados, e são esses campos que orientam as experiências. Nesse sentido, a informação fornecida pelos instrumentos poderia apoiar o desenvolvimento do modelo de metadados. Contudo, para que isso fosse possível seria necessário aceder aos grupos de investigação e vários laboratórios onde os instrumentos estão localizados, sendo isto identificado como uma tarefa «muito difícil», uma vez que seria necessário «pedir autorizações a várias pessoas ligadas aos laboratórios».

Da análise de conteúdo nem sempre foi possível identificar ao que o autor se referia, como por exemplo ao identificar determinado instrumento nem sempre foi possível compreender qual a sua função, tendo em conta a falta de conhecimento do curador no domínio em causa. Resultou assim na apresentação do descritor *Instrumento* e de instrumentos com as funções específicas como *Instrumento de radiação de luz UV* ou *Instrumento de medição do pH.* O investigador afirmou que a especificação da função do instrumento tornava mais fácil a descrição. Verificou-se ainda a importância de registar todos os instrumentos utilizados ao longo da investigação, sobretudo para alguns instrumentos, onde pode ser relevante recolher informação sobre calibrações efetuadas, quem produziu o instrumento, entre outras referências. Neste caso não foi possível definir um descritor que captasse esses valores, apesar de nos artigos analisados serem, por vezes, feitas referências a este tipo de especificidades.

Para além disso, importa registar toda a informação sobre as variáveis que possam ter influência nos resultados, como métodos, técnicas, temperaturas registadas.

Discussão

A hipótese que sustenta o desenvolvimento deste estudo assenta na possibilidade de a análise de conteúdo apresentar características que melhoram o trabalho do curador de dados. O objetivo passa por compreender se é possível que o curador, sem conhecimento específico do domínio, consiga de forma autónoma, identificar descritores fundamentais para a descrição dos dados, de modo a que estes sirvam como ponto de partida para o estabelecimento da interação entre curador e investigador.

Com base na avaliação efetuada, de um total de 60 descritores propostos ao investigador, 53 foram identificados como indo ao encontro das suas necessidades de descrição. A percentagem de sucesso foi de aproximadamente 88% fornecendo uma primeira validação da relevância da abordagem. Desta sessão de avaliação junto do investigador é possível fazer algumas considerações sobre a abordagem.

Primeiro, a facilidade em comunicar com o investigador uma vez que o trabalho proativo do curador faz com que este esteja preparado para compreender melhor as necessidades de descrição do domínio e ainda, torna-o capaz de apresentar exemplos práticos de descritores que poderão ser relevantes. Por sua vez, o investigador é estimulado a uma maior participação, pois vê materializado, através dos exemplos de descritores, o que é necessário para descrever os seus dados.

Segundo, a agenda dos investigadores é dedicada essencialmente ao processo de investigação que é prioritário, assim acredita-se que a análise de conteúdo se apresenta como uma solução viável uma vez que é possível desenvolver um modelo de metadados ajustado às necessidades dos domínios num período de tempo aproximado a uma semana, sem que o investigador necessite despender muito do seu tempo.

Apesar dos benefícios apontados a abordagem tem aspetos a melhorar. O investigador refere que o modelo apresentado contém mais elementos do que necessita para a descrição da sua experiência. Isso evidencia que é possível uma menor exaustividade da análise de conteúdo. Com um menor número de descritores propostos dá-se a oportunidade ao investigador para propor descritores que considere relevantes, sem que seja influenciado por uma lista demasiado extensa.

Constata-se também que em alguns casos o investigador tem preferência por descritores mais genéricos como *Método de análise* em vez de *Método de análise de polifenóis*, enquanto noutros prefere a especificação do descritor, *Instrumento de medição do pH* em vez de *Instrumento*. Estas escolhas levantam algumas questões: 1) será que um nível mais genérico é suficiente para captar toda a informação necessária para a descrição do contexto experimental, sem que ocorram perdas de informação; 2) a especificação em demasia dos descritores será uma solução melhor, uma vez que orienta o investigador, porém pode limitar o seu contributo aquando da interação com o curador. É importante garantir que o investigador descreve toda a informação relevante orientando para tal com descritores mais específicos. O investigador vê o descritor e sabe exatamente o que descrever, limitando assim a perda de informação que pode ocorrer quando o descritor é mais genérico. O investigador pode não se lembrar de anotar valores importantes pois tem de incluir muita informação num só campo. Por sua vez, é necessário considerar que a especificidade em demasia pode dar origem a modelos de metadados demasiado extensos e consecutivamente impraticáveis.

Este trabalho é apenas o início de um longo percurso. Com base no modelo de metadados desenvolvido é possível partir para a construção de ontologias, que podem ser aplicadas em sistemas que apoiem a gestão de dados. A coleção de novos casos de uso é importante para compreender possíveis similaridades entre domínios, com novos casos de estudo a aprendizagem e possível reutilização dos descritores pelo curador poderá tornar a tarefa de análise de conteúdo mais ágil.

Este trabalho vai continuar com novos casos e avaliação junto dos investigadores, onde se tentará verificar se é possível atingir resultados semelhantes em condições diferentes, considerando a possibilidade de se conseguir generalizar requisitos comuns a diversos domínios.

Conclusões

Conclui-se, com base nos resultados obtidos, que a análise de conteúdo é uma abordagem com características relevantes para a otimização do trabalho desenvolvido pelo curador de dados. O curador sem conhecimentos no domínio consegue de uma forma prática identificar conceitos fundamentais à descrição de conjuntos de dados.

A proatividade do curador garante um input inicial que torna a comunicação entre o curador e o investigador, que a avaliação sugere ser mais eficaz. Essa facilidade estimula o investigador que através de exemplos práticos tem mais facilidade em compreender em que consiste a gestão de dados. Este argumento pode ser justificado à luz da intervenção efetuada pelo investigador, que a determinado momento da avaliação, sugere que se o curador tivesse acesso à informação predefinida dos instrumentos o trabalho poderia ser facilitado. Esta sugestão acontece no instante em que o curador perceciona que a informação que se lhe estava a pedir com os metadados era semelhante à que tinha de registar nos instrumentos.

Contudo, existe ainda espaço para que a abordagem seja melhorada, futuramente a abordagem pode ser efetuada de forma menos exaustiva tornando a tarefa mais rápida e permitindo que a

intervenção do investigador seja maior. Outro aspeto interessante seria a avaliação dos resultados junto de mais investigadores.

Referências bibliográficas

BORGMAN, Christine - Advances in information Science: The Conundrum of Sharing Research Data. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY** . ISSN 19335954. 63(6). 2012. 1059–1078. doi: 10.1002/asi.

CHAO, Tiffany C. - Identifying Indicators of Description for Research Data from Scientific Journal Publications. 2014.

EUROPEAN COMMISSION - **Directorate-General for Research & Innovation. H2020 Programme, Guidelines on FAIR Data Management in Horizon 2020** [Em linha] Disponível em WWW:<URL:http://ec.europa.eu/research/%0Aparticipants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.>.2016.

EUROPEAN COMMISSION - **H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020**. 2017.

THANOS, Costantino - Research Data Reusability: Conceptual Foundations, Barriers and Enabling Technologies. **Publications**. ISSN 2304-6775. 5:1 (2017) 19. doi: 10.3390/publications5010002.

UK DATA ARCHIVE - **Managing and sharing data best practice for researchers**. 2011.

WHITE, Hollie C. - Descriptive Metadata for Scientific Data Repositories: A Comparison of Information Scientist and Scientist Organizing Behaviors. **Journal of Library Metadata**. ISSN 19375034. 14:1 (2014) 24–51. doi: 10.1080/19386389.2014.891896.

WILJES, Cord; CIMIANO, Philipp - Linked Data for the Natural Sciences: Two Use Cases in Chemistry and Biology. **Proceedings of the Workshop on the Semantic Publishing (SePublica 2012)**. 2012.

WILLIS CRAIG, GREENBERG JANE, White Hollie - Analysis and Synthesis of Metadata Goals for Scientific Data. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY**. 63(8) . 2012. 1505–1520. doi: DOI: 10.1002/asi.