I CONFERÊNCIA INTERNACIONAL
DE GESTÃO DA INFORMAÇÃO E ARQUIVOS

Dataset Identifiers para a Comunicação em Ciência: Exploração e Propostas

Luis Corujo

Universidade de Lisboa, Faculdade de Letras, Centro de Estudos Clássicos

luiscorujo@campus.ul.pt

Palavras-chave: Dados Abertos, *Datasets*; *Dataset Identifiers,* Metainformação

Introdução

Os estudos sobre a Ciência desenvolvidos no âmbito de iniciativas da Ciência Aberta (CA) permitiram elucidar problemas como o inflacionamento e enviesamento de *rankings* de pesquisas de publicações (Young, 2008) e as limitações em termos de reprodutibilidade devido aos dados não serem distribuídos com as publicações (Kraker, 2011). A CA pretende abrir o processo de investigação, ao defender a reprodução dos resultados da investigação, uma metodologia de investigação transparente, e assim aumentar o impacto social do investigador e economizar tempo e dinheiro dos investigadores e das instituições de investigação (RIN/NESTA, 2010). Recorre para tal a instrumentos e/ou componentes denominados de acesso aberto (uma forma de tornar os resultados de investigação disponíveis), dados abertos (como uma maneira de publicar os dados em bruto), código aberto (como uma maneira de dar acesso a protótipos de investigação), a metodologia aberta (como uma forma de partilha dos detalhes metodológicos do estudo e das ferramentas utilizadas para a recolha e análise de dados), investigação reprodutível aberta (o ato de praticar CA para permitir a reprodutibilidade independente dos resultados da investigação), avaliação pelos pares aberta (garantia de qualidade transparente e verificável através de revisão aberta), recursos educacionais abertos (usar materiais livres e abertos para a educação e o ensino universitário) (Kraker, 2011; Pontika, 2015; Kasberger, 2013).

Paralelamente, a utilização de tecnologia digital de informação no meio científico tem provocado alterações a nível experimental e teórico, dando origem a um "dilúvio de dados" que, de acordo com alguns autores, é representativo de um novo paradigma científico - "*data intensive*" – conotado com a *e-Science*, e que tem como fim um mundo em que não só a literatura científica está acessível em linha, mas também os dados científicos, e em que ambos são interoperáveis (Hey et al., 2009, p. xxx). Tal visão, ligada à exploração de conjuntos de dados (*datasets)* capturados por

instrumentos eletrónicos, gerados por simulações em computador e sensores em rede, requer um conjunto de ferramentas e tecnologias de apoio à colaboração científica e cooperação entre instituições de investigação, para análise e mineração de dados (*data mining*), para visualização e exploração de dados, e para a comunicação e disseminação académica (Hey et al., 2009b, pp. xviii-xxx).

Os benefícios advindos da comunicação e tratamento intensivo destes volumes imensos de conjuntos de dados, publicações e materiais de apoio em constante mutação e evolução prendem-se não só com o fim da visão das ciências, separadas como torres de marfim e silos de informação, mas também com a troca de aprendizagens, experiências e melhores práticas com origem no entrosamento inter/trans/multi- disciplinar. Também a promoção do acelerar de descobertas e o destacar de novas ligações, sugerindo ligações previamente imprevistas, impulsionam o desenvolvimento científico. (Ginsparg, 2009; Lynch 2009).

Nesta lógica impõe-se a abordagem da questão da disponibilização dos dados científicos integrando-a no movimento do Acesso Aberto e dos Dados Científicos Abertos. O objetivo é identificar o contexto motivacional para a identificação e citação de dados científicos, e, imbricado nessa motivação, compreender o contributo da análise quantitativa aliada à citação de dados científicos para as métricas da comunicação científica. Pretende-se conhecer as propostas tecnológicas utilizadas para a identificação e citação desses dados, e identificar a meta-informação, documentos e registos a ter em conta para a gestão, preservação a longo prazo e acesso dos dados científicos, na perspetiva da comunidade arquivística. A motivação deste estudo está ligada à urgência na criação de serviços de apoio à gestão de dados científicos no seio dos serviços de informação e documentação das instituições de investigação e ensino superior, para cumprimento das políticas de dados dos organismos financiadores da ciência. Está em causa a promoção de uma cultura de dados abertos que incentive a partilha, identificação e a citação dos dados de investigação, o acesso às fontes de informação e o apoio a novas formas de partilhar ciência (GT-BES, 2015; Lopes, 2016).

Em termos de metodologia e dados utilizados, procedeu-se a uma revisão bibliográfica relativa aos dados científicos abertos usando como fontes para pesquisa bases de dados da EBSCO e B-On, e subsequente análise dos artefactos, explorando as perceções dos autores relativamente aos dados científicos abertos e a sua citação.

Dados Científicos e Dados Abertos

Os dados foram sempre a pedra basilar da ciência, pois não é possível replicar resultados experimentais, executar atividades de investigação observacional, ou testar afirmações sem eles, sendo assim prova do conjunto do conhecimento científico. Pode-se afirmar que a ciência é construída através da coleta, análise, publicação, reanálise crítica e reutilização dos dados (Socha, 2013; Molloy 2011). São produtos essenciais decorrentes dos e úteis para os princípios científicos básicos, como a reprodutibilidade e a transparência (Socha, 2013), o que os torna em ativos

intelectuais de primeira importância, passível de revisão por pares, avaliação de qualidade e reutilização (Heidorn, 2011) embora sem permissão explícita (Murray-Rust, 2008). Os Dados tornaram-se um foco crítico para a comunicação científica, gestão da informação e política de investigação (Borgman, 2012), sendo disponibilizados em maior quantidade em linha, e o acesso a grandes coleções de dados é cada vez mais solicitado para fins de educação, ciência, política e comércio (Altman, 2013).

Entende-se por **dados digitais** não só as manifestações digitais como texto, som, imagem, modelos, jogos e simulações, mas também formas de dados e bases de dados cuja usabilidade requer apoio do sistema intermediário (*hardware* e *software*), tais como os variados tipos de dados laboratoriais de espectrografia, sequenciação genética, microscopia de eletrões, os dados observacionais como os dados de teledeteção, espaciais e socioeconómicos, e outras formas de dados gerados e/ou compilados por humanos ou máquinas (Uhlir & Cohen, 2011, *apud* Borgman, 2011, p. 1061). Um **conjunto de dados (*Dataset*)** é definido como uma coleção identificável de dados (ISO 19115, 2003) que pode conter um ou mais ficheiros de dados em formato idêntico, com as mesmas variáveis e especificações. Tratam-se muitas vezes de dados não-textuais, em formato digital, que constituem dados brutos a ser trabalhados e dos quais resultam dados progressivamente mais refinados (Buckland, 2011). Este pode conter conteúdos originais ou ser produto derivado de outros dados e/ou versões (Peng, 2016). A sua organização deve permitir a pesquisa e recuperação e/ou processamento e reorganização (Socha, 2013).

Para que a sociedade beneficie amplamente desses dados é necessário que sejam disponibilizados de uma forma útil e aberta, ou seja, sem preço ou barreiras. Molloy (2011) refere que quanto maior for a quantidade de dados abertos, maior será o nível de transparência e reprodutibilidade e, logo, a eficiência do processo cientifico.

As Declarações emanadas de iniciativas da comunidade científica como a *Budapest Open Access Initiative* (2002), a *Bethesda Statement on Open Access Publishing* (2003) e a *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* (2003) demonstram as potencialidades das tecnologias da informação para fornecimento de acesso aberto à informação científica, englobando literatura revista e avaliada pelos pares, e os dados da investigação. Discute-se como pôr em prática a partilha em acesso aberto de forma a garantir que a reutilização dessa informação permita a produção de trabalhos dela derivados. Esta questão ganha ainda mais relevância quando está em causa a investigação fruto do investimento público de entidades nacionais e internacionais, dando mais consistência à obrigação destes dados serem disponibilizados publicamente em infraestruturas eletrónicas que garantam a sua acessibilidade, usabilidade e reutilização, e ainda para garantir a verificação de boas práticas científicas e o retorno do investimento público (OECD, 2007; Comissão Europeia, 2012). Isto dará origem a políticas de financiamento da investigação que requerem que os resultados e os dados recolhidos que lhes deram origem sejam disponibilizados abertamente.

Os *Panton Principles* consideram dados científicos abertos aqueles que estão livremente disponíveis na Internet, permitindo que qualquer utilizador os descarregue, copie, analise, reprocesse, os transforme em aplicações informáticas,

ou os utilize para qualquer outra finalidade sem quaisquer outras barreiras financeiras, legais ou técnicas que não sejam as derivadas do acesso à própria internet. Tal significa que todos os dados relacionados com a publicação científica devem estar explicitamente no domínio público (Murray-Rust et al., 2010). Os Dados abertos são o próximo passo lógico depois de acesso aberto: uma vez que os resultados são publicados de forma agregada, são necessários os dados em bruto para tornar estes resultados reproduzíveis. Além de possibilitar a tarefa de reproduzir os resultados da investigação, os dados abertos permitem a avaliação das diferentes hipóteses sobre o mesmo conjunto de dados. Nesse sentido, os dados abertos podem ser entendidos como a reutilização e reciclagem de dados da investigação sem ter que gastar recursos com outro ciclo de coleta de dados (Kraker, 2011). Os dados abertos também permitem a agregação de dados de vários estudos para ganhar novos conhecimentos (Murray-Rust, 2008). Para Borgman (2011) a motivação para fazer avançar a investigação ou para servir o bem público pode estar subjacente às políticas de partilha de dados de agências de financiamento, revistas científicas e outras partes interessadas. Os argumentos variam também pelo fato de servirem os interesses dos investigadores que produzem os dados ou os interesses dos potenciais utilizadores desses dados. Os incentivos para os investigadores partilharem dados dependem de muitos fatores, incluindo não só o argumento da partilha, mas os tipos de dados, os fins para os quais foram recolhidos, as abordagens para a coleta de dados e processamento, as preocupações com a potencial utilização indevida ou má interpretação, recursos para documentação de dados, e os meios para a sua curadoria e divulgação. Borgman (2011) resume as razões que levam à partilha dos dados a quatro tipos de argumentos que variam ao longo das dimensões de motivação para a partilha de dados científicos e os interesses por trás dessa partilha de dados: para reproduzir ou verificar a investigação, para tornar os resultados da investigação financiada publicamente disponíveis ao público, para permitir que outras pessoas elaborem novas questões acerca dos dados existentes, e para avançar o estado da investigação e inovação. O primeiro argumento é o mais forte do ponto de vista da investigação, uma vez que as questões de reprodutibilidade estão profundamente entrelaçadas com a epistemologia do campo da investigação. O segundo e terceiro argumentos são mais movidos pelo interesse público e são apresentados na perspetiva daqueles que desejam usar os dados produzidos por outras partes. O quarto argumento, que também beneficia o público, está enquadrado no interesse dos produtores de dados, e serve a investigação, inovação e a academia.

Encontram-se exemplos de partilha de dados científicos no âmbito da criação do *World Data System* (WDS) do *International Council of Science* (ICSU) aquando do Ano Internacional da Geofísica em 1957/8, no Projeto do Genoma Humano em 1996 e a investigação relativa à bactéria *Escherichia coli*, esta já com a publicação da sequência genética na plataforma *GitHub* e com licenciamento da *Creative Commons* (Pampel, 2014).

A promoção da partilha de dados requer a identificação e resolução das barreiras que influenciam os cientistas na partilha dos dados da sua investigação, sendo que as principais incluem a falta de recompensa ou de crédito para a partilha, as questões legais, o uso indevido dos dados e má interpretação dos dados, o controlo sobre a propriedade intelectual, bem como a necessidade de restringir o acesso ou extirpar dados de identificação de seres humanos ou de espécies ameaçadas de extinção, incompatibilidade de tipos de dados e o trabalho para os documentar em formas reutilizáveis, tempo insuficiente e falta de financiamento derivado do faco de poucas entidades financiadores promoverem a inovação pela reutilização dos dados (Kuipers e Hoeven, 2009; Tenopir et al., 2011; Borgman, 2011). Centrando estas barreiras no âmbito das infraestruturas dedicadas, Graaf e Waaijers (2011) consideram ser necessária a concessão de incentivos para estimular a partilha de dados, a intensificação da formação e educação dos cientistas e dos prestadores de serviços ligados ao tratamento e manuseio dos dados, o desenvolvimento de redes de infraestruturas para armazenamento perante e confiável dos dados de investigação, e consciencializar para as questões do seu financiamento a longo prazo.

Para garantir o sucesso da partilha de dados há que criar condições para que os cientistas considerem positivo a disponibilização dos dados da sua investigação, o que, para Pampel e Dallmeier-Tiessen (2014) implica que a partilha de dados se torne um critério de base para o sistema de reputação científico existente, nomeadamente através de estratégias ligadas à publicação de dados de investigação como um objeto de informação independente num repositório de dados científicos, como documentação textual na forma de um "artigo de dados", ou como elemento de enriquecimento de um artigo ou "publicação aprimorada". Tais estratégias recorrem para a tecnologia que permite a descrição e endereçamento persistente dos *datasets*.

Citação de dados

A citação de dados, de acordo com o Australian National Data Service (ANDS) pode ser definida como a prática de fornecer uma referência aos dados de forma similar às citações que os investigadores fazem com as referências bibliográficas dos recursos publicados. No caso das citações bibliográficas dizem respeito a uma referência estruturada e formal relativa a um outro trabalho académico, podendo ocorrer várias vezes ao longo do texto, com a referência bibliográfica completa para o trabalho a ser incluída numa lista de referências, muitas vezes após o final do texto principal. Estas citações incluem indicações, como a identificação das páginas do documento citado a que diz respeito a referência. Esta prática está de tal forma estabelecida que a maioria das editoras e das organizações de investigação (e financiamento) têm manuais de instruções que fornecem os detalhes de como deve ser fornecida a informação e como as referências devem ser estruturadas (cf. APA, 2010; NP405-1, 1994). A citação de dados é uma referência aos dados para efeitos de atribuição de crédito e facilitação do acesso aos dados. No entanto, como os conjuntos de dados se vão tornando maiores e mais complexos, tornou-se muitas vezes impossível publicá-los como parte de um artigo, pelo que o escrutínio das

afirmações feitas nas publicações científicas requer que a ligação entre os dados e a publicação seja mantida. O surgimento da referenciação de trabalhos digitais permitiu uma maior variedade nas formas de citação dessas obras, pelo que o uso corrente deixa em aberto a questão da terminologia usada para descrever as referências dos dados com maior nível de granularidade, incluindo os subconjuntos de observações, variáveis, ou outros componentes e subconjuntos de um conjunto de dados maior. Estas referências granulares são muitas vezes necessárias nos trabalhos para descrever os elementos que o veiculam na forma de tabela de dados, gravura, ou análise e são análogas aos elementos de citação utilizados para referenciar as páginas nos artigos científicos (Socha, 2013).

Altman (2013) refere a existência de quatro fases de desenvolvimento na área da citação de dados entre 1977 e a atualidade. A primeira fase focou-se no papel da citação para facilitar a descrição e recuperação de informação, introduzindo os princípios que os dados nos arquivos devem ser descritos como trabalhos e não simples suportes, utilizando autor, título e versão. A segunda fase estendeu as citações para suporte ao acesso e identificação persistente dos dados, devendo as citações ser diretamente acionáveis na web; a terceira fase focou-se na utilização de citações para verificação e reprodutibilidade, iniciando a tendência para a ampla integração com o ecossistema de publicação; a quarta e atual fase centra-se na integração no ecossistema académico de investigação e publicação, incluindo a integração da citação de dados normalizada dentro das publicações, catálogos, redes de ferramentas e grandes sistemas de atribuição.

A avaliação das práticas e análise da literatura sobre as práticas de citação permitem identificar os princípios para a citação de dados (Socha, 2013) que serão formalizados em 2014 na *Joint Declaration of Data Citation Principles* (Data Citation Synthesis Group, 2014)

Tabela 1- Joint Declaration of Data Citation Principles (Data Citation Synthesis Group, 2014)

1	Importância	Os dados devem ser considerados produtos de investigação legítimos e citáveis. Deve ser dada a mesma importância às citações de dados que é dada às citações de outros objetos de investigação, tais como publicações, em termos de registo académico
2	Crédito e Atribuição	As citações de dados devem facilitar a dar o devido crédito académico e atribuição normativa e legal para todos os contribuidores dos dados, reconhecendo não pode ser aplicável um único estilo ou mecanismo de atribuição a todos os dados.
3	Evidência	Na literatura académica, quando uma afirmação se baseia em dados, devem ser citados os dados correspondentes
4	Identifícação unívoca	A citação de dados deve incluir um método de identificação persistente acionável por máquina,

		globalmente exclusivo, e amplamente utilizado por uma comunidade
5	Acesso	As citações de dados devem facilitar o acesso aos próprios dados e a qualquer meta-informação, documentação, código e outros materiais associados necessários para os seres humanos e máquinas fazerem uso informado dos dados referenciados.
6	Persistência	Os identificadores únicos e meta-informação que descrevem os dados, e sua organização, devem ser persistentes - mesmo para além do tempo de vida dos dados que descrevem.
7	Especificidade e verificabilidade	As citações de dados devem facilitar a identificação, acesso e verificação dos dados específicos que suportam uma afirmação. As citações de dados ou a sua meta-informação devem incluir informação suficiente sobre a proveniência e a estabilidade (fixidez) para facilitar a verificação de que a iteração temporal, versão e / ou porção granular de dados posteriormente recuperados é a mesma que foi originalmente citada.
8	Interoperabilidade e flexibilidade	Os métodos de citação de dados devem ser suficientemente flexíveis para acomodar as diferentes práticas das várias comunidades, mas não deve diferir tanto ao ponto de comprometer a interoperabilidade das práticas de citação dados entre as comunidades.

Os estudos acerca da reutilização de dados de Piwowar (2013) permitem concluir a existência de um conjunto de benefícios ligados à utilização citação de dados:

- em termos de reutilização de dados, em que os artigos com conjuntos de dados disponíveis podem ser usados de maneiras que os estudos sem dados não podem, podendo resultar num maior número de citações;
- em termos de credibilidade, na medida em que a credibilidade dos resultados da investigação pode ser maior para artigos com dados disponíveis, levando a que esses documentos sejam preferencialmente escolhidos como citações de fundo ou como base de investigação adicional;
- em termos de maior visibilidade, uma vez que há maior probabilidade de outros investigadores encontrarem artigos com os dados disponíveis, seja através de uma ligação direta a partir dos dados ou indiretamente através de promoção cruzada, como o caso das hiperligações de um repositório de dados a um artigo poderem aumentar o *ranking* de pesquisa do trabalho de investigação; em termos de visualização antecipada, quando os dados são disponibilizados antes do artigo ser publicado, pode obter algumas citações mais cedo do que seria possível de outra forma, por causa da consciência acelerada dos métodos, resultados, *etc*.;

- em termos de viés de escolha, uma vez que os autores podem ser mais propensos a publicar dados para artigos que julgam ser o seu trabalho de melhor qualidade, derivado do seu carácter orgulhoso ou confiante dos resultados.

Socha (2013) aborda ainda os benefícios e barreiras à adoção de boas práticas de citação de dados, identificando-os ao nível dos criadores de dados, para os administradores das Universidades e instituições de investigação, para os centros de dados (*datacenters*), para as organizações de financiamento, para os editores, para os investigadores e comunidades de investigação, e para a sociedade em geral, adicionando ainda as barreiras económicas e financeiras.

Tal como a citação tradicional de publicações, a citação de dados pode apoiar as métricas da produção académica e permitir o crédito devido a todas as partes que contribuem para a investigação, além de fornecer atributos detalhados, facilitando o acesso futuro, e promovendo a colaboração e investigação cruzadas (Wickett, 2012). Além disso, a citação de dados tem o potencial de ir além da citação tradicional de publicações ao fazer a ligação entre os resultados da investigação científica publicada e os conjuntos de dados relevantes – necessário para manter a integridade do todo científico, facilitando assim a reutilização de dados e a descoberta orientada por dados (Wickett, 2012; Socha, 2013). Alguns financiadores de investigação começaram a exigir que os dados da investigação financiada publicamente sejam depositados em vários centros de dados. À medida que essas práticas se vão estabelecendo, a capacidade de detetar, localizar, obter, e compreender os dados de investigações anteriores ficará circunscrita à capacidade de ter uma descrição suficiente desses dados: a citação (Socha, 2013).

Apesar destes benefícios potenciais, as práticas atuais de citação de dados estão longe de estarem amadurecidas. Exemplo disso é o fato destas práticas estarem intimamente baseadas nas necessidades dos investigadores que trabalham com determinados tipos de dados, o que constitui uma barreira para a reutilização de dados em novos contextos ou descobertas orientadas por dados, por falta de adoção noutras áreas científicas (Wickett, 2012).

Na generalidade, as diferentes partes interessadas geralmente têm objetivos comuns no âmbito da investigação, apesar de poderem defender e promover diferentes perspetivas e interesses. A fim de implementar de forma abrangente boas práticas de citação de dados em algumas disciplinas, será necessário cativar a maioria, se não a totalidade destas partes interessadas, requerendo para tal um enfoque especializado e um esforço substancial e sustentado (Socha, 2013).

As barreiras identificadas passam também pelas questões ligadas aos dados: a *big data*, os dados com estruturas complexas, dados estruturados e os dados em formatos em mudança (Altman, 2013).

Altman (2013) e Socha (2013) partilham da ideia de que a implementação de sistemas de citação de dados ainda se debate com três categorias de desafios: de

proveniência, que inclui a cadeia de propriedade de um objeto, e a história das suas transformações, e cujos modelos têm fortes implicações na forma como a citação de dados está integrada no fluxo de curadoria de dados; de identidade, cujas teorias incluem a definição dos dados, a identidade dos dados e como definir as relações de equivalência e de derivação, e a granularidade e estrutura dos dados., sendo que as teorias de dados têm fortes implicações na determinação do que deve ser citado; de atribuição, que desempenha um papel fundamental no incentivo às citações, sendo que os seus modelos têm fortes implicações na determinação da apresentação das citações de dados.

Outras questões de fundo passam por aspetos como o domínio da investigação (reprodutibilidade da investigação, diminuição do viés de publicação e/ou inexistência de viés de resultados, reutilização da investigação, as revisões, sumários e meta-análise da investigação, e a análise interdisciplinar), as políticas de Ciência (robustez e fiabilidade da ciência, transparência e políticas baseadas na ciência, e o impacto do financiamento na ciência) e a cienciometria (o impacto direto na partilha, replicação e retração de dados, os padrões de citação dos dados e artigos relacionados, as métricas de citação para os dados e outras métricas de impacto, e a construção de um novo mapas da ciência (Socha, 2013).

No caso específico da utilização de métricas de impacto académico nos conjuntos de dados Wickett (2012) afirma a necessidade de coordenação entre os editores, que expõem a citação de dados para a indexação, e os produtores das métricas de impacto, que terão que desenvolver e refinar as métricas de impacto acadêmico dos dados da investigação para incorporar a citação de dados de investigação. Ingwersen (2011;2014) aponta o potencial da utilização dos conjuntos de dados e a sua citação para o desenvolvimento de estudos cienciométricos, informétricos, webométricos (onde se inclina para incluir os altmétricos relativos ao uso dos média sociais), cibermétricos e bibliométricos. Para o efeito desenvolve um Índice de Uso de Dados com o objetivo dar visibilidade à utilização de conjuntos de dado e o devido reconhecimento aos seus criadores, gestores e editores, incentivando os editores de aumentar o volume de descoberta, mobilidade e publicação de dados de grande qualidade, aumentar o uso de dados primários na tomada de decisões com suporte científico, e aumentar o comportamento da citação formal de conjuntos de dados na investigação. Os catorze indicadores definidos incluem os registos pesquisados, a frequência de descarregamento (dos dados da Internet), o número de revistos, o número de pesquisas, o número de descarregamentos, o número de *datasets*, a densidade de pesquisa, a densidade de descarregamento, o impacto de uso, o impacto de interesse, o rácio de uso, o balanço de uso, a pontuação de uso e a pontuação de interesse (Ingwersen, 2011). Considera-se que os estudos baseados nestas métricas podem permitir verificar se existem as similitudes de características entre os artigos de jornais científicos e os *datasets*, e entre *datasets* de diferentes áreas científicas, para além de servir de complemento à monitorização da

investigação baseada na citação tradicional e para a avaliação institucional (Ingwersen, 2014).

Tipos/Metodologias de Citação de dados

Reconhecem-se vários métodos que os investigadores utilizam atualmente para a citação de dados. Wickett (2012) apresenta uma série de abordagens que incluem a utilização de uma secção de reconhecimento numa publicação a dar crédito aos fornecedores dos dados, a citação de um trabalho de investigação onde o conjunto de dados é introduzido e usado, e a citação de um artigo publicado que incida sobre os dados, caso tal publicação estiver disponível. No entanto estas abordagens sofrem de problemas ligados à falta de suporte às métricas da produção académica, por não refletirem a natureza dinâmica de muitos dos conjuntos de dados, uma vez que podem ser atualizados frequentemente. Os trabalhos publicados são estáticos e podem ficar desatualizados, resultando numa barreira ligada à aplicação de modelos conceptuais de gestão de informação projetados para lidar com objetos estáticos, como artigos, na gestão de conjuntos de dados que estão sempre a ser modificados e acrescentados. Uma solução para essas questões é citar o próprio conjunto de dados bem descrito (por exemplo, com meta-informação descritiva e de proveniência, bem como hiperligações para estudos relacionados) e desenvolver acordos para normas de citação de dados que usem identificadores únicos e persistentes para conjuntos de dados. Silvello (2015), baseando-se em Altman (2013) e do Data Citation Synthesis Group (2014), propõe quatro requisitos principais que devem ser cumpridos pelas metodologias de citação de dados:

1 - fornecer uma descrição dos dados, com a finalidade de dar crédito académico e atribuições normativas e legais para os criadores e curadores dos dados. Mesmo que ainda não hajam regras que definam que meta-informação de descrição compõe uma citação completa, mas há um certo consenso sobre o conjunto mínimo exigido, que deve conter: autor, título, data e local (ou seja, uma referência persistente para os dados citados);

2 - identificar univocamente o objeto citado e a meta-informação associada;

3 - permitir citações de dados com vários níveis de granularidade (para permitir citar um conjunto de dados como um todo, uma unidade única ou um subconjunto de dados);

4 - produzir referências legíveis por humanos e máquinas.

Pampel (2014), baseando-se em Dallmeier-Tiessen (2012) propõe três estratégias de publicação dos dados da investigação já testadas: 1 - como um objeto de informação independente num repositório dados da investigação, usado há muito tempo nas ciências biológicas com o uso de repositórios de dados, tais como o *GenBank* (Benson, 2012), 2 - como o enriquecimento de um artigo, apelidado de "publicação enriquecida", como são exemplo os indicado na Tabela 2 e que usam também a ligação de artigos e dados (Woutersen-Windhouwer et al. 2009), sendo o

objetivo a construção e manutenção de um ambiente técnico que relacione todos os objetos de informação relevantes em torno de um artigo para que seja criado um espaço de conhecimento onde os dados de investigação que servem de base ao artigo podem ser livremente acessíveis; 3 - como uma documentação textual, também chamado de artigo de dados e cuja a intenção é descrever os dados, em vez de relatar um estudo de investigação (Chavan & Penev 2011), como são exemplos os *American Geophysical Union* (AGU) *Journals* e os *Journals Ecological Archives of the Ecological Society of America* (ESA). Atualmente já inclui os chamados jornais de dados, como o *Open Access journal Earth System Science Data* (ESSD), que existe desde 2008 (Pfeiffenberger, 2011), sendo que os conjuntos de dados são publicados num repositório confiável, e tanto estes como as publicações descritiva têm um apontador persistente com base num identificador de objeto digital (DOI, 2016), que também facilita a citação de dados. Estes DOI são constituídos por um código alfanumérico único e persistente compatível com a web, que aponta para um recurso (por exemplo, um conjunto de dados) a ser preservado a longo prazo (ou seja, durante várias gerações de hardware e de software) (Socha, 2013)

Graças a este procedimento, que foi desenvolvido no âmbito do projeto *Publication and Citation of Scientific Primary Data* (STD-DOI) (Klump, 2006) e expandida pelo DataCite (Brase, 2011; 2013; 2015), é possível ligar as publicações e os dados subjacentes. Este procedimento também suporta a visibilidade dos dados. Algumas editoras, têm, por exemplo, integrado dados de investigação de acesso livre nas suas plataformas (Reilly, 2011). Uma série de revistas científicas de dados também surgiu mesmo período (ver tabela 2). De notar que a criação revistas de dados só tem viabilidade com o livre acesso aos dados, metainformação e o texto publicado correspondente, para permitir a reutilização e aumentar o impacto nas métricas de citação.

Tabela 2- Exemplos de Publicações "Enriquecidas" e de artigos revistas de dados respetivas editoras (baseado em Pampel, 2014)

Publicações	Editoras
Atomic Data and Nuclear Data Tables	Elsevier
Nuclear Data Sheets	Elsevier
Biodiversity Data Journal	Pensoft Publishers
Dataset Papers in Biology	Hindawi Publishing Corporation
Dataset Papers in Chemistry	Hindawi Publishing Corporation
Dataset Papers in Ecology	Hindawi Publishing Corporation
Dataset Papers in Geosciences	Hindawi Publishing Corporation
Dataset Papers in Materials Science	Hindawi Publishing Corporation
Dataset Papers in Medicine	Hindawi Publishing Corporation
Dataset Papers in Nanotechnology	Hindawi Publishing Corporation
Dataset Papers in Neuroscience	Hindawi Publishing Corporation
Dataset Papers in Pharmacology	Hindawi Publishing Corporation

Dataset Papers in Physics	Hindawi Publishing Corporation
Earth System Science Data - ESSD	Copernicus Publications
Geoscience Data Journal	Wiley
GigaScience	BioMed Central
Open Network Biology	BioMed Central
Open Archaeology Data	Ubiquity Press

Accomazzi (Accomazzi et al., 2011), no âmbito da utilização de identificadores persistentes, propõe três modelos de citação de dados: 1 - Citar dados como artigos, atribuindo metainformação básica aos produtos de dados (autor, título), um identificador persistente (DOI ou outro), e incluí-los nas Referências bibliográficas, juntamente com todos os outros documentos, tal com o modelo recomendado para citação de dados no *Dryad* (Vision, 2010) e no *DataVerse* (Crosas, 2011); entre outros; 2 - citar dados como websites: descobrir qual o URI (supostamente persistente) e mencioná-lo no artigo como uma referência em linha ou uma nota de rodapé, tal como o modelo adotado nos arquivos da NASA via ADS (Accomazzi, 2011); e 3 - Ter uma seção de referências de dados nos artigos, tal como a lista de referências bibliográficas, essa seção iria listar de maneira inequívoca (e com formatação normalizada) todos os produtos de dados utilizados no estudo que originou o artigo, facilitando a identificação das citações de dados para os editores, curadores e agregadores.

A proposta do consórcio internacional *DataCite*, uma iniciativa e uma norma de metainformação, inclui universidades, instituições de investigação, agências de gestão de dados e entidades governamentais. Muitas dessas instituições estão agrupadas na Europa através da Biblioteca Nacional Alemã de Ciência e Tecnologia (TIB), com presença significativa na América do Norte, Austrália (através do *Australian National Data Service*), e na Ásia. O seu intuito é prestar apoio aos investigadores (ajudando-os a encontrar, identificar e citar dados de investigação e outros objetos de investigação de maneira fiável), aos centros de centros (fornecendo identificadores persistentes para conjuntos de dados, fluxos de trabalho e normas para publicação de dados), e aos editores de revistas científicas, permitindo a ligação dos artigos científicos aos dados/objetos subjacentes (Adamich, 2016) .Partindo das propostas deste consórcio, Harvey (2015) apresenta três métodos: 1 - utilizando os recursos do *Handle System* suportados pela *Corporation for National Research Initiatives* (CNRI, 2015) e que é, também a tecnologia subjacente por trás do sistema *DOI*. Normalmente, o registo *handle* de um *DOI* consiste um valor *URL*, para o qual o navegador é redirecionado quando o *DOI* é resolvido por um agente como um servidor *proxy http://doi.org/* ou *http: //hdl.handle .net/.* Este *URL* normalmente aponta para uma página legível por humanos. 2 - Método 2: usando o resolvedor de conteúdo e API de formatos *DataCite*: A API de depósito de metainformação *DataCite* (MDS, 2016) inclui um recurso de formatos que serve para os tipos MIME poderem ser associados a URLs como pares de chave de valor. Em vez de redirecionar para a

página de destino habitual, um DOI pode, então, direcionar para esses URLs alternativos através da negociação de conteúdo. Isto congrega tecnologia desenvolvido através do *Crosscite* (2016) com o Resolvedor de conteúdo *DataCite*. Para usar este recurso uma aplicação pode resolver um determinado DOI, enquanto especifica os formatos de ficheiro aceitáveis numa lista de tipos MIME no cabeçalho *Accept* da solicitação HTTP. A resolução devolve assim um que corresponde a um dos formatos solicitados, desde que esse ficheiro de dados tenha sido registado com a identificação do seu tipo usando a API de formatos *DataCite*. O resolvedor de conteúdo também apresenta as URLs registradas com a API de formatos através de ligações em HTML. O resolvedor de conteúdo *DataCite* apresenta limitações quando um conjunto de ficheiros tem mais do que um ficheiro do mesmo tipo MIME. Uma forma de evitar conflitos entre tipos de MIME é atribuir um DOI para cada ficheiro, seguindo os princípios de granularidade funcional; 3 – Método 3: Mapas de Recursos OAI-ORE expostos através da metainformação *DataCite*: Algumas das limitações do método anterior podem ser superadas pela exposição da estrutura de diretórios do conjunto de ficheiros publicados de uma forma localizável e legível por máquinas. ORE (20008) (*Object Reuse and Exchange*) é uma norma da *Open Archives Initiative* (OAI) para descrever agregações de recursos da web através de documentos referidos como mapas de recursos. Estes podem ser serializados em vários formatos, incluindo Atom (usado para feeds RSS), RDF ou RDF-a (para declarar RDF triplos). O servidor de repositório cria automaticamente um mapa de recursos *Atom OAI-ORE* para cada objeto de repositório, bem como um ficheiro de metainformação METS (2016) (*metadata encoding and transmission standard*), que é uma norma alternativa que também fornece metainformação estrutural adequada. Estes ficheiros podem ser tornados detetáveis ao incluir as respetivas localizações como identificadores relacionados na metainformação *DataCite* do objeto de repositório. A metainformação é mapeada internamente no esquema *Dublin Core*. A metainformação *DataCite* de um determinado DOI pode ser recuperado através da negociação de conteúdo e, a partir desta, a metainformação ORE o ou METS, se for encontrada, pode ser recuperada e processada para devolver o URL de um ficheiro específico. A exposição do URL do Mapa de recursos OAI-ORE através da metainformação DOI também fornece uma solução para o problema da detetabilidade dos mapas de recursos ORE.

Este estudo considera que as soluções apresentadas no seio das propostas do Consórcio *Datacite* são as que mais das respostas pretendidas em termos de metodologias de citação de dados.

Metainformação e informação associada a *datasets*

As Citações e a meta-informação são interdependentes. As citações académicas "tradicionais" são utilizadas para a atribuição do crédito, verificação e localização dos artigos, livros ou sítios web citados, existindo, porém, vários estilos definidos por convenções bibliográficas de cada área científica. No caso das citações de dados, trata-se de referência aos dados para efeito de atribuição de crédito e facilitação do acesso aos mesmos, incorporando normalmente um número limitado de elementos de meta-

informação, como um identificador persistente, o título descritivo e informação de estabilidade (fixidez) para verificação da proveniência. Os objetos de dados descritos pela citação são geralmente detetáveis por esta meta-informação de citação. Adicionalmente, o objeto de dados está geralmente associado a meta-informação mais completa do que a fornecida pela citação, pelo que esta é útil para, indiretamente, aceder a um conjunto maior de meta-informação (Socha, 2013). Esta meta-informação adicional pode incluir informação descritiva e de representação sobre o objeto, para uma maior transparência e inteligibilidade; informação sobre os sistemas intermediários utilizados para gerar o objeto, para o reforço da transparência e reprodutibilidade; informação sobre os procedimentos desenvolvidos para garantir a qualidade do produto para aumentar a confiabilidade dos dados; e informação sobre como obter ficheiros de dados e usá-los para aumentar o potencial de descoberta e usabilidade. A captura e transmissão de meta-informação de uma forma consistente irá melhorar a interoperabilidade, e facilitar a comparação dos derivados dos dados por parte dos utilizadores com o fim de determinar quais consideram mais adequados para as aplicações pretendidas (Peng, 2016). Isto liga-se intimamente à qualidade científica que se refere à fiabilidade, precisão, validade e adequação do produto às aplicações pretendidas (Ramapriyan et al., 2015).

Neste âmbito Socha (2013) apresentava a proposta de elementos de meta-informação para citação de dados do *Digital Curation Centre* (DCC) (Ball, 2015), que consistem no autor, título, editor, data de publicação, tipo de recurso, edição, versão, o nome específico e URI, dados de verificação, o identificador, e a localização. Socha (2013) indicava ainda práticas incompletas e lacunas referentes aos elementos de meta-informação e citação de dados, ligados às questões de granularidade, controlo de versões, facilitação da reutilização, a identificação de pequenos contributos e atribuição de mérito científico-académico, que o consórcio *Datacite* parecer pretender resolver com a indicação de cinco atributos obrigatórios (Datacite, 2016; Adamich, 2016) (Identificador, criador, título, editor, ano de publicação) e treze opcionais (assunto, contribuidores, data, língua, tipo de recurso, identificador alternativo, identificador relacionado, tamanho, formato, versão, direitos, descrição, e geolocalização). Este estudo aventa ainda a proposta de utilização para este âmbito do esquema de *Meta-informação para a Interoperabilidade* (MIP), desenvolvido pela Direção-Geral de Arquivo (DGARQ, 2012), na medida em que pretende prover a interoperabilidade entre organismos ao nível da utilização, gestão e acesso a recursos informativos, entendidos como qualquer objeto informacional, contendo ou veiculando informação de diferente natureza, e que detenha identidade dentro do universo de discurso, sendo aplicável a qualquer recurso produzido ou detido por uma organização e independentemente do suporte ou formato em que é produzido, seja ele um documento de arquivo, bibliográfico, museológico, um serviço, uma referência ,um sitio web, etc. Considera-se que esta proposta tem cabimento por dar resposta às lacunas e problemas anteriormente identificados e não estar limitado a nenhuma área científica.

Outra proposta que este trabalho apresenta, prende-se com um aspeto que não surge muito aprofundada na bibliografia científica, e que diz respeito à necessidade de implementar procedimentos de gestão e preservação de informação de arquivo aos registos e documentos que fornecem e reforçam valor probatório e fiabilidade aos conjuntos de dados científicos. Considera-se vital a proposta do Conselho

Internacional de Arquivos (Arovelius, 2010) para a gestão e preservação de dados científicos e documentos de arquivo a eles associados (não dos dados científicos em si!), uma vez que propõe uma avaliação arquivística dessa documentação e registos, independentemente do suporte e do formato, fornecendo tabelas de seleção documental consentâneas com as várias fases/atividades do processo de investigação em ciência, a saber: Planeamento, Coleta de dados, Análise, Avaliação, Reportar dos resultados, Relatório financeiros, e Arquivo. Este documento também fornece exemplos ligados a Universidades e laboratórios dos EUA, Suécia e Brasil.

Conclusão

Verifica-se a existência de um campo de estudo ligado à citação de dados, derivado da comunicação em ciência principalmente por intermédio dos aspetos da ciência aberta relacionados com o acesso aberto e dados abertos. Pretende diagnosticar o uso da citação de dados no seio das ciências, promover a citação de dados e desenvolver soluções para os desafios e barreiras que identifica nas várias áreas científicas e das partes interessadas. Entre essas soluções contam-se o desenvolvimento de ferramentas tecnológicas para a citação de dados de forma estável, unívoca e persistente, e o enquadramento contextual dos dados, do ponto de vista sincrónico e diacrónico, de todas as suas iterações e estados, através da meta-informação, documentação e registos de ações exercidas no sistema onde se encontram os dados.

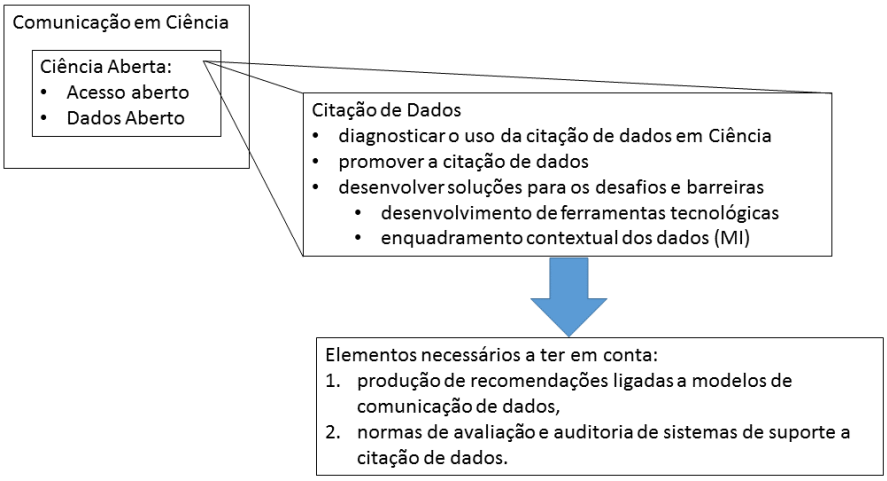

Figura 1 - Esquema de sistematização no âmbito do estudo das citações de dados

Constata-se a existência de numerosos estudos sobre identificação de artigos e publicações científicas e formas de os identificar e citar (ex. DOI), e também sobre como identificar os autores, os investigadores, os académicos (ex. ORCID), mas existem relativamente poucos estudos de fundo sobre a identificação dos conjuntos de dados científicos. Este estudo pretendeu em conclusão fomentar o uso de dados aberto para criar novo conhecimento e novos serviços (economia do conhecimento),

evidenciando os elementos necessários a ter em conta na produção de recomendações ligadas a modelos de comunicação de dados, e normas de avaliação e auditoria de sistemas de suporte a citação de dados.

No que se refere ao dinamismo e preservação dos *datasets*, evidenciou-se a questão de guardar ou não cada iteração dos dados. Trata-se de um aspeto pouco tratado, na medida em que se fala apenas em termos do dinamismo e mutabilidade/alteração e acrescentos nos dados. Mas se estes servem também de prova ao que está publicado, sendo necessário garantias de segurança e confiança nesses dados e saber os conteúdos de dados em determinada data, quem fez o acrescento/alteração e quando. Tal informação deve ficar no registo de auditoria do sistema onde estão alojados os dados.

Tal também demonstra a relação entre a questão dos *datasets* e comunicação em ciência, identificada no âmbito das métricas e também na questão do acesso aberto, pelo acesso aos dados e pela comunicação nos artigos e jornais. Isto porque a publicação de dados de investigação é requerida cada vez mais pelas editoras das revistas científicas que, para garantir maior transparência, querem agora não só os artigos de investigação, mas também os dados.

A promoção da citação de dados irá fomentar um sistema de comunicação científica que permita a identificação, recuperação e atribuição de dados de investigação, em que os repositórios que publiquem os dados devem estipular como condições de reutilização a meta-informação e citações mandatadas. A identificação dos dados disponíveis através da literatura publicada será sempre uma estratégia de referenciação problemática até que a citação de dados consistente proporcione um mecanismo de recuperação normalizado, resultando esta expetativa no incentivo à partilha de dados, à promoção da investigação secundária, e à melhoria do ritmo e da qualidade do intercâmbio na comunidade académica (Mooney, 2012)
A ligação entre os conjuntos de dados e as publicações por intermédio das citações de dados é um passo importante para tornar a investigação mais reprodutível e transparente (Piwowar, 2013). O desempenho científico deve, no futuro, ser avaliado com um "fator de partilha" que não só considere a frequência de citação na comunidade científica, mas também dê importância à implementação da partilha de informação e conhecimento para o bem da sociedade (Dallmeier-Tiessen, 2014).

Finalmente, este estudo apresenta o estado da arte em termos de tecnologia/ferramentas de citação de dados, nos quais se enfatiza a proposta do consórcio *Datacite.* Adicionalmente, propõe um esquema de metainformação para identificação dos conjuntos de dados – o MIP (*MetaInformação para a InteroPerabilidade*) da atual DGLAB – que se considera cumprir com as necessidades de granularidade, controlo de versões, facilitação da reutilização, identificação de pequenos contributos e atribuição de mérito científico-académico. Expõe ainda a necessidade de implementar procedimentos de gestão e preservação de informação de arquivo referente aos dados científicos, para garantir o valor probatório e a fiabilidade dos conjuntos de dados científicos.

Referências Bibliográficas

Accomazzi, A. (2011), "Linking literature and data: Status report and future efforts", *Future Professional Communication in Astronomy II*, Springer, pp. 135–142.

Accomazzi, A., Derriere, S., Biemesderfer, C. and Gray, N. (2011), "Why don't we already have an Integrated Framework for the Publication and Preservation of all Data Products?", *arXiv Preprint arXiv:1112.1688*, available at: http://arxiv.org/abs/1112.1688 [Acedido em 27 de julho de 2017].

Adamich, T. (2016), "Accessing Scholarly Research Datasets using DataCite, ORCID, and CASRAI.", *Technicalities*, Vol. 36 No. 3, pp. 18–21.

Altman, M. and Crosas, M. (2013), "The evolution of data citation: From principles to implementation.", *IAssist Quarterly*, Vol. 37 No. 1–4, pp. 62–70.

American Psychological Association [APA]. (2010). *Publication Manual of the American Psychological Association*. 6ª Ed. Washington D.C. APA

AROVELIUS, R. and others. (2010), "Management and preservation of scientific records and data", *International Council on Archives: Paris*, pp. 11–12.

Australian National Data Service [ANDS]. [s/d]. *Data citation awareness guide.* http://www.ands.org.au/guides/data-citation-awareness . [Acedido em 27 de julho de 2017]

Ball, A. & Duke, M. (2015). 'How to Cite Datasets and Link to Publications'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides [Acedido em 27 de julho de 2017]

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J. and Sayers, E.W. (2012), "GenBank", *Nucleic Acids Research*, Vol. 40 No. D1, pp. D48–D53.

"Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities". (2003). Berlim. Max Planck Geselschaft, http://openaccess.mpg.de/Berlin-Declaration. [Acedido em 27 de julho de 2017]

Brown, P. et al. (2003). "Bethesda Statement on Open Access Publishing". http://www.earlham.edu/~peters/fos/bethesda.htm [Acedido em 27 de julho de 2017]

Borgman, C.L. (2011), "The Conundrum of Sharing Research Data", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 6, pp. 1059–1078.

Borgman, C. (2012), "Why Are the Attribution and Citation of Scientific Data Important?", in Uhlir, P.E. and others. (2012), *For Attribution–Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*, National Academies Press, pp 1-8

Brase, J. and Farquhar, A. (2011), "Access to Research Data", *D-Lib Magazine*, Vol. 17 No. 1/2, available at: http://doi.org/10.1045/january2011-brase. [Acedido em 27 de julho de 2017]

Brase, J. (2013), "DataCite and linked data", *JLIS. It*, Vol. 4 No. 1, p. 365.

Brase, J., Lautenschlager, M. and Sens, I. (2015), "The Tenth Anniversary of Assigning DOI Names to Scientific Data and a Five Year History of DataCite", *D-Lib Magazine*, Vol. 21 No. 1/2, available at: http://doi.org/10.1045/january2015-brase. [Acedido em 27 de julho de 2017]

Buckland, M. (2011), "Data Management as Bibliography.", *Bulletin of the American Society for Information Science & Technology*, Vol. 37 No. 6, pp. 34–37.

Budapest Open Access Initiative. (2002). Declaration after the Open Society Institute meeting in Budapest December 1-2 2001. Budapest: Open Society Institute. http://www.soros.org/openaccess/read.shtml. [Acedido em 27 de julho de 2017]

Chavan, V. and Penev, L. (2011), "The data paper: a mechanism to incentivize data publishing in biodiversity science", BMC Bioinformatics, Vol. 12 No. 15, p. 1.

Comissão Europeia. (2012). *Commission recommendation on access to and preservation of scientific information*. Bruxelas. Comissão Europeia.

Corporation for National Research Initiatives (CNRI) (2015) http://www.cnri.reston.va.us/. [Acedido em 27 de julho de 2017]

Crosas, M. (2011), "The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data", *D-Lib Magazine*, Vol. 17 No. 1/2, available at http://doi.org/10.1045/january2011-crosas. [Acedido em 27 de julho de 2017]

Crosscite (2016) : http://www.crosscite.org/cn. [Acedido em 27 de julho de 2017]

Dallmeier-Tiessen, S., Darby, R., Gitmans, K., Lambert, S., Suhonen, J. and Wilson, M. (2012), "Compilation of results on drivers and barriers and new opportunities", available at: http://epic.awi.de/31394/ [Acedido em 27 de julho de 2017].

Data Citation Synthesis Group (2014), "Joint Declaration of Data Citation Principles". Martone M. (ed.) San Diego CA: FORCE11. Disponível em https://www.force11.org/group/joint-declaration-data-citation-principles-final. [Acedido em 27 de julho de 2017].

DataCite, (2016) www.datacite.org [Acedido em 27 de julho de 2017].

DataCite Metadata Store (MDS), (2015) http://mds.datacite.org/. [Acedido em 27 de julho de 2017].

DGARQ (2012), "Metainformação para a Interoperabilidade (MIP)", disponível em http://arquivos.dglab.gov.pt/wp-content/uploads/sites/16/2013/10/MIP_v1-0c.pdf [Acedido em 27 de julho de 2017]

Digital object identifier (DOI) system (2016) http://www.doi.org/ [Acedido em 27 de julho de 2017]

Ginsparg, P. (2009), "Text in a Data-centric World", *The Fourth Paradigm: Data-Intensive Scientific Discovery.*, Microsoft Research, Redmond, Washington, pp. 185–191.

Graaf, M., van der; Waaijers, L. (2011). *A surfboard for riding the wave. towards a four country action programme on research data.* Disponível em: http://www.umic.pt/images/stories/publicacoes5/KE_Surfboard_Riding_the_Wave_Screen.pdf [Acedido em 27 de julho de 2017]

GT-BES. (2015). *Recomendações para as bibliotecas de ensino superior*. Lisboa: Associação Portuguesa de Bibliotecários, Arquivistas e Documentalistas; 2015.

Harvey, M.J., Mason, N.J., McLean, A. and Rzepa, H.S. (2015), "Standards-based metadata procedures for retrieving data for display or mining utilizing persistent (data-DOI) identifiers", *Journal of Cheminformatics*, Vol. 7 No. 1, available at:http://doi.org/10.1186/s13321-015-0081-7. [Acedido em 27 de julho de 2017]

Heidorn, P.B. (2011), "The Emerging Role of Libraries in Data Curation and E-science.", *Journal of Library Administration*, Vol. 51 No. 7/8, pp. 662–672.

Hey, A.J.G., Tansley, S. and Tolle, K.M. (Eds.). (2009a), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, Washington.

Hey, A.J.G., Tansley, S. and Tolle, K.M. (2009b), "Jim Gray on eScience: A Transformed Scientific Method", *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, Washington, pp. xvi–xxxi.

Ingwersen, P. and Chavan, V. (2011), "Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure", *BMC Bioinformatics*, Vol. 12 No. 15, p. 1.

Ingwersen, P. (2014), "Scientific Datasets: Informetric Characteristics and Social Utility Metrics for Biodiversity Data Sources", in Chen, C. and Larsen, R. (Eds.), *Library and Information Sciences*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 107–117.

ISO 19115 (2003): Geographic Information — Metadata. I*nternational Organization of Standards*. Version: ISO 19115:2003

Kasberger, S. (2013), "Open Science", *openscienceASAP*, 8 August, disponível em: http://openscienceasap.org/open-science/ [Acedido em 27 de julho de 2017].

Kraker, P., Leony, D., Reinhardt, W. and Beham, G. (2011), "The case for an open science in technology enhanced learning", *International Journal of Technology Enhanced Learning*, Vol. 3 No. 6, pp. 643–654.

Kuipers, T. & Hoeven, J., van der. (2009). *Insight into digital preservation of research output in europe. survey report*, Disponível em: http://libereurope.eu/wp-content/uploads/2010/01/PARSE.Insight.-Deliverable-D3.4-Survey-Report.-of-research-output-Europe-Title-of-Deliverable-Survey-Report.pdf. [Acedido em 27 de julho de 2017]

Lopes, C.A. (2016), "Bibliotecas de ensino superior: Novas e saudáveis tendências", *XII Jornadas APDIS*, pp. 1–10.

Lynch, C. (2009), "Jim Gray's fourth paradigm and the construction of the scientific record", *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, Washington, pp. 177–183.

Metadata Exchange and Transmission Standard (METS). (2016) disponível em http:// www.loc.gov/standards/mets/ [Acedido em 27 de julho de 2017].

Molloy, J.C. (2011), "The Open Knowledge Foundation: Open Data Means Better Science.", *PLoS Biology*, Vol. 9 No. 12, pp. 1–4.

Mooney, H. and Newton, M. (2012), "The Anatomy of a Data Citation: Discovery, Reuse, and Credit", *Journal of Librarianship and Scholarly Communication*, Vol. 1 No. 1, available at:http://doi.org/10.7710/2162-3309.1035.

Murray-Rust, P. (2008), "Open Data in Science.", *Serials Review*, Vol. 34 No. 1, pp. 52–64.

Murray-Rust, Peter; Neylon, Cameron; Pollock, Rufus; Wilbanks, John. (2010). *Panton Principles: Principles for Open Data in Science.* Disponível em: http://pantonprinciples.org/. [Acedido em 27 de julho de 2017]

NP 405-1. (1994). Informação e Documentação. Referências bibliográficas: documentos impressos. Lisboa: IPQ.

Object Reuse and Exchange (OAI-ORE) (2008), http://www.openarchives.org/ore/1.0/discovery [Acedido em 27 de julho de 2017]

Pampel, H. and Dallmeier-Tiessen, S. (2014), "Open Research Data: From Vision to Practice", Opening Science, Springer, disponível em:

http://book.openingscience.org/vision/open_research_data.html [Acedido em 27 de julho de 2017]

OCDE (2007). *Principles and Guidelines for Access to Research Data from Public Funding.* OECD.

Peng, G., Ritchey, N.A., Casey, K.S., Kearns, E.J., Privette, J.L., Saunders, D., Jones, P., et al. (2016), "Scientific Stewardship in the Open Data and Big Data Era — Roles and Responsibilities of Stewards and Other Major Product Stakeholders", *D-Lib Magazine*, Vol. 22 No. 5/6, available at:http://doi.org/10.1045/may2016-peng.

Pfeiffenberger, H. and Carlson, D. (2011), "'Earth System Science Data' (ESSD) - A Peer Reviewed Journal for Publication of Data", *D-Lib Magazine*, Vol. 17 No. 1/2, available at:http://doi.org/10.1045/january2011-pfeiffenberger.

Piwowar, H.A. and Vision, T.J. (2013), "Data reuse and the open data citation advantage", *PeerJ*, Vol. 1, p. e175.

Pontika, N., Knoth, P., Cancellieri, M. and Pearce, S. (2015), "Fostering open science to research using a taxonomy and an eLearning portal", ACM Press, pp. 1–8.

Poole, A. (2015), "How has your science data grown? Digital curation and the human factor: a critical literature review.", *Archival Science*, Vol. 15 No. 2, pp. 101–139.

Ramapriyan, H., D. Moroni, e G. Peng (2015): Improving information quality for Earth Science data and products — An overview. #IN14A-01. *AGU 2015 Fall Meeting*, San Francisco, CA, USA

Reilly, S., Schallier, W., Schrimpf, S., Smit, E. and Wilkinson, M. (2011), "Report on integration of data and publications", available at: http://epic.awi.de/31397/ [Acedido em 27 de julho de 2017]

RIN/NESTA. (2010), "Open to all? Case studies of openness in research", available at: http://www.rin.ac.uk/system/files/attachments/NESTA-RIN_Open_Science_V01_0.pdf [Acedido em 27 de julho de 2017]

Socha, Y.M. (Ed.). (2013), "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data", *Data Science Journal*, Vol. 12, pp. 1–75.

Silvello, G. (2015), "A Methodology for Citing Linked Open Data Subsets", *D-Lib Magazine*, Vol. 21 No. 1/2

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., et al. (2011), "Data Sharing by Scientists: Practices and Perceptions", *PLoS ONE*, Vol. 6 No. 6, p. e21101.

Vision, T.J. (2010), "Open Data and the Social Contract of Scientific Publishing", *BioScience*, Vol. 60 No. 5, pp. 330–331.

Wickett, K.M., Xiao, H. and Thomer, A. (2012), "The RDAP12 Summit: Challenges and Opportunities for Data Management.", *Bulletin of the American Society for Information Science & Technology*, Vol. 38 No. 5, pp. 14–19.

Woutersen-Windhouwer, S. (Ed.). (2009), *Enhanced Publications: Linking Publications and Research Data in Digital Repositories*, Amsterdam Univ. Press, Amsterdam.

Young, N.S., Ioannidis, J.P. and Al-Ubaydli, O. (2008), "Why current publication practices may distort science. The market for exchange of scientific information: the winner's curse, artificial scarcity, and uncertainty in biomedical publication", *PLoS Medicine*, available at: http://files.figshare.com/454023/Text_S1.pdf [Acedido em 27 de julho de 2017]

Anexo I - Elementos e subelementos do esquema MIP (Metainformação para a Interoperabilidade)

Elemento	Subelemento	Sub-subelemento
1. Designação	1. Título	
1. Designação	1.1. Alternativo	
1. Designação	1.2. Atribuído	
2. Identificador	2.1. Tipo de Identificador	
2. Identificador	2.2. Id de recurso	
2. Identificador	2.3. Código de classificação	
2. Identificador	2.4. Versão	
3. Produtor	3.1 Sector	
3. Produtor	3.2 Id do Produtor	
3. Produtor	3.3 Designação do Produtor,	
3. Produtor	3.4 Tipo de Produtor	
4. Assunto		
5. Descrição		
5. Descrição	5.1. Idioma	
5. Descrição	5.2. Descritores	
6. Editor	6.1 Id Editor	
6. Editor	6.2. Nome Editor	
6. Editor	6.3. Tipo de editor	
7. Colaborador	7.1 Sector,	
7. Colaborador	7.2 Id do colaborador	
7. Colaborador	7.3 Designação do colaborador	
7. Colaborador	7.4 Tipo de colaborador,	
8. Data	8.1 Data/hora de criação	
8. Data	8.2 data/hora de registo	
8. Data	8.3 data/hora de aquisição,	
8. Data	8.4 data/hora de disponibilidade	
8. Data	8.5 data/hora de abertura	
8. Data	8.6 data/hora encerramento	
9. Tipo de Recurso		
10. Formato	10.1. Formato de dados	
10. Formato	10.2. Dimensão	
10. Formato	10.3. Suportes	
11. Relação	11.1. Identificador de recurso relacionado	
11. Relação	11.2. Tipo de relação	11.2.1. Tem versão de
11. Relação	11.2. Tipo de relação	11.2.2. É versão de
11. Relação	11.2. Tipo de relação	11.2.3. Tem parte de
11. Relação	11.2. Tipo de relação	11.2.4. É parte de

Elemento	Subelemento	Sub-subelemento
11. Relação	11.2. Tipo de relação	11.2.5. É formato de
11. Relação	11.2. Tipo de relação	11.2.6. É referenciado por
11. Relação	11.2. Tipo de relação	11.2.7. Referencia
11. Relação	11.2. Tipo de relação	11.2.8. É substituído por
11. Relação	11.2. Tipo de relação	11.2.9. Substitui
11. Relação	11.2. Tipo de relação	11.2.10. É requerido por
11. Relação	11.2. Tipo de relação	11.2.11. Requer
11. Relação	11.3. Descrição da rel	
12. Cobertura	12.1. Espacial	
12. Cobertura	12.2. Temporal	
13. Acessibilidade	13.1. Classificação de Segurança	
13. Acessibilidade	13.2. Estatuto de utilização	
13. Acessibilidade	13.3. Condições de publicação	
13. Acessibilidade	13.4. Encriptação	
13. Acessibilidade	13.5. Autenticação de assinatura electrónica	
13. Acessibilidade	13.6. Permissões de acesso de grupo	
13. Acessibilidade	13.7. Lista de circulação	
13. Acessibilidade	13.8. Copyright	
14. Destinatário	14.1 Sector	
14. Destinatário	14.2 Id do Destinatário	
14. Destinatário	14.3 Designação do destinatário,	
14. Destinatário	14.4 Tipo de destinatário	
15. Disponibilidade	15.1. Custo/preço	
15. Disponibilidade	15.2. Localização	
16. Avaliação	16.1. Prazo de Conservação	
16. Avaliação	16.2. Destino Final	
17. Agregação		