Schöch, Christof. Acesso aberto para as máquinas. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 21, jan./jun. 2020, p. 99-115. Tradução de Winfried Nöth.

Recebido em: 19 mar. 2020
Aprovado em: 22 abr. 2020

dx.doi.org/
10.23925/1984-3585.2020i21p99-115

Acesso aberto para as máquinas

Christof Schöch[1]

Tradução do alemão[2]: Winfried Nöth

Resumo: Muitos debates sobre o Acesso Livre giram atualmente em torno de modelos de financiamento adequados, perdendo de vista o fato de que os formatos de publicação também devem ser considerados como fundamentalmente mais abertos, ou seja, para além do formato PDF como a reencarnação digital do livro impresso, para que o potencial das tecnologias digitais de produção de conhecimento possa ser utilizado da melhor forma possível. Por conseguinte, o presente artigo trata da questão das publicações científicas como dados (mas não da publicação de dados de investigação). A exigência central do documento é trabalhar no sentido da substituição do formato PDF e do desenvolvimento e utilização de formatos de dados abertos e padronizados para publicações científicas que satisfaçam os princípios FAIR. O artigo argumenta que, para além dos metadados, das palavras-chave, da estrutura do texto e das referências bibliográficas, o conteúdo do texto é de particular importância: são necessários modelos e soluções técnicas para a forma como as declarações centrais de uma publicação científica podem ser incorporadas na própria publicação de uma forma legível por máquina. Para todos estes aspectos, formatos semiestruturados amplamente utilizados como XML (por exemplo TEI, JATS ou RDF) e JSON (como em BibJSON, semelhante também a BibTex), bem como o princípio da Web Semântica com a utilização de Dados Abertos Vinculados desempenham um papel importante.

Palavras-chave: Dados Abertos Vinculados. Acesso aberto. Publicação científica. XML. BibTex.

---

1 Professor de Humanidades Digitais na Universidade de Trier, Alemanha, codiretor do Centro Trier de Humanidades Digitais, presidente da COST Action Distant Reading for European Literary History e da Associação de Humanidades Digitais nos países de língua alemã (DHd) dh.uni-trier.de.

2 de Schöch (2020).

Open access to machines

Abstract: The debates about Open Access currently revolve around the question of suitable financing models, losing sight of the fact that publication formats must also be thought of as fundamentally more open. That means looking beyond the PDF format as the digital reincarnation of the printed book, if the full potential of digital technology to produce knowledge is to be exploited in the best possible way. This paper therefore deals with the question of scientific publications as data (but not with the publication of research data). The central demand of this paper is to work towards a replacement of the PDF format and the development and use of open, standardised data formats for scientific publications that meet the FAIR principles. The paper argues that, in addition to metadata, keywords, text structure and bibliographic references, text content is of particular importance. Models and technical solutions are needed for how central statements of a scientific publication can be embedded in the publication itself in a machine-readable form. For all these aspects an important role can be played by the use of semi-structured formats such as XML (for example TEI, JATS or RDF) and JSON (as in BibJSON, similar to BibTex), as well as the semantic web with the use of Linked Open Data.

Keywords: Linked Open Data. Open Access. Scientific publishing. XML. BibTex.

Debates atuais sobre acesso aberto

O debate sobre o Acesso Livre nas ciências e humanidades mudou consideravelmente nos últimos anos. Mesmo que a prática fique muitas vezes um pouco atrás das convicções e da medida em que os servidores de *preprints* são utilizados, a proporção de revistas de acesso aberto ou a disponibilidade gratuita dos anais de conferências varia, visto que em grandes partes do sistema científico, pelo menos para artigos de revistas e artigos de conferências, já não há qualquer dúvida de que a publicação em acesso aberto faz sentido e é cientificamente apropriada (ver open-access.net/en/open-access-in-individual-disciplines para as diversas disciplinas). Um dos argumentos mais comuns é que os resultados da investigação com financiamento público devem também estar disponíveis ao público em sentido lato; e que o progresso científico pode ser mais bem promovido pela livre disponibilidade de publicações científicas em todo o mundo.

O debate atual é, portanto, menos sobre o porquê do que sobre o como. Nisso, prevalece um enfoque claro sobre a questão dos modelos de financiamento adequados. A questão central é como é que os custos de publicação e divulgação, por um lado, e os custos de desenvolvimento e disponibilidade de infraestruturas de publicação a longo prazo, por outro, podem ser financiados se isso já não for feito através de subscrições, como é atualmente o caso. Entre os modelos atualmente discutidos e praticados estão os chamados Encargos de Processamento de Artigos (isto é, as taxas de publicação a serem cobradas dos autores, suas instituições ou patrocinadores do projeto), grandes acordos de leitura e publicação (tais como os contratos com as principais editoras em nível nacional que o consórcio DEAL está tentando concluir) ou novos modelos de financiamento coletivo (tais como o modelo de adesão à Biblioteca Aberta de Humanidades; <open-access.net/en/information-on-open-access/business-models>; SPEICHER *et al.* 2018). Como pode ser feita uma relocação dos orçamentos de assinatura por bibliotecas para a promoção de editoras e iniciativas que publicam em Acesso Livre? Como se pode evitar que a atual injustiça de acesso (apenas aqueles que podem pagar são autorizados a ler os resultados científicos de outros) seja meramente substituída por uma

injustiça de publicação (apenas aqueles que podem pagar são autorizados a publicar resultados científicos) (Piron, 2020; Pooley, 2020)? Esta última questão em particular tem também uma forte dimensão internacional e é, portanto, também de importância para a política de desenvolvimento.

A solução da questão do financiamento é sem dúvida de grande importância, não apenas, mas especialmente para as Humanidades. Entretanto, a presente contribuição terá que se concentrar num outro aspecto da questão – que atualmente está sendo negligenciado nos debates intensivos sobre o financiamento –, a saber, a questão dos formatos de publicação adequados para a ciência. Na prática, o arquivo PDF, que é o equivalente digital do livro impresso e do artigo de periódico, domina claramente. Três características explicam a aceitação e o sucesso deste formato: (1) correspondência direta entre a versão impressa e a versão digital até o layout; (2) preservação da paginação e, portanto, possibilidade de continuar as práticas de citação familiar; (3) a aparente imutabilidade e, portanto, confiabilidade de um arquivo PDF. No entanto, exceto para distribuição e leitura individual, este formato só é adequado até certo ponto (apesar de algumas extensões como PDF/A para arquivamento e PDF etiquetado para melhor acessibilidade). Neste contexto, o Grupo de Trabalho de Publicação Digital da Associação Alemã *Digital Humanities* recomenda, por exemplo: "não use o PDF como o formato primário de publicação (camada de codificação), mas sim, se for o caso, como um formato de leitura derivado" (DHd-AG, 2020). Tanto para o arquivamento a longo prazo como para a avaliação computadorizada de coleções maiores de publicações, outros formatos têm claras vantagens.

Publicações científicas como dados

Assim que se trata menos de um punhado de publicações da sempre crescente literatura de pesquisa, ou seja, se trata não mais apenas de uma questão de leitura, mas de usar as publicações como base de dados para uma análise quantitativa, os numerosos pontos fracos do formato PDF se tornam aparentes. Se os princípios FAIR (*Findable*, *Accessible*, *Interoperable*, *Re-useable*) forem aplicados às publicações científicas como dados em vez de à publicação de dados de pesquisa, torna-se rapidamente claro o quão desastroso é a prática atualmente dominante de publicar exclusivamente como arquivos PDF (Wilkinson; Dumontier; Mons, 2016). É verdade que tais contribuições podem ser encontradas (são *findable*) através de identificadores e metadados persistentes, que estão fre-

quentemente disponíveis atualmente. Publicadas em Acesso Livre, elas também são acessíveis (*accessible*) sem grandes obstáculos financeiros ou técnicos. Porém, elas também são só parcialmente interoperáveis e reutilizáveis (*interoperable* e *re-useable*). Assim, o texto em um arquivo PDF pode ser extraído, mas praticamente sem nenhuma informação estrutural essencial. A separação entre título corrido, texto principal e anotações é, na melhor das hipóteses, apenas acessível indiretamente, através de referências tipográficas ou outros padrões. Dentro do texto principal, não é possível distinguir de forma confiável entre diferentes seções de texto (por exemplo, resumo, introdução, parte analítica ou interpretativa, resultados, ou mesmo entre texto principal e citações em bloco). Também informações semânticas não estão explicitamente disponíveis, porque entidades (pessoas, obras, organizações) ou conceitos (termos técnicos ou abstratos) não podem ser abordados especificamente dentro do texto. Da mesma forma, não é possível pesquisar especificamente autores ou autoras, editores, títulos, datas de publicação ou organizadores dentro dos dados bibliográficos.

Nessas circunstâncias, as publicações científicas não podem fazer valer plenamente o seu potencial. Eles só podem fazê-lo se, além de serem publicados digitalmente e com livre acesso, também estiverem disponíveis em formatos estruturados e semanticamente enriquecidos. As estratégias de publicação correspondentes, que tanto quanto possível produzem não apenas publicações legíveis por humanos, mas também por máquinas, têm sido discutidas há dez anos (com base na ideia da Web Semântica) sob o título de *Semantic Publishing* (Shotton, 2009). Semelhante ao caso da construção e publicação de dados nas ciências humanas, grandes quantidades de publicações científicas são úteis, mas ainda melhor são publicações semanticamente e estruturalmente enriquecidas, que assim se tornam conjuntos de dados.

Alguns cenários de aplicação relevantes neste contexto estão resumidamente delineados a seguir. Por exemplo, a análise linguística da linguagem científica está interessada nas propriedades linguísticas de textos de diferentes disciplinas ou de diferentes tipos de literatura científica. Ela poderia lidar com dados de publicação anotados estruturalmente, mas também com vocabulário, estilística e padrões de argumentação de seções funcionalmente diferentes de textos científicos (tais como introdução, parte principal ou conclusão). A pesquisa quantitativa sobre a história de uma disciplina, por exemplo, poderia ganhar uma base empírica muito mais ampla através de análises detalhadas e em larga escala de redes

de citação baseadas em bibliografias estruturadas. E a pesquisa baseada em um *corpus* de textos sobre um determinado autor, obra ou problema, que já se beneficia de resumos e palavras-chave, poderia funcionar muito mais precisa e extensivamente, se as entidades essenciais no texto fossem anotadas e as declarações centrais de todas as publicações fossem legíveis por máquina e automaticamente vinculáveis a uma rede de declarações.

Os requisitos essenciais para as possibilidades que tais publicações científicas legíveis por máquina devem oferecer podem ser resumidos da seguinte forma:

1. Codificação estruturada e padronizada de metadados relacionados a documentos (incluindo informações bibliográficas; palavras-chave; licenças; identificadores persistentes, tais como DOIs);

2. Codificação explícita das estruturas de texto (entre outros, texto principal versus anotações; introdução, parte principal, conclusão; se necessário, dados, hipóteses, métodos, resultados; texto do autor versus citações);

3. Codificação estruturada de referências bibliográficas (incluindo informações bibliográficas, incluindo identificadores persistentes, como DOIs para literatura de pesquisa e, quando apropriado, fontes primárias);

4. Rotulagem legível por máquina das entidades (atores, organizações, lugares, horários) e conceitos em uma contribuição (resumo, termos técnicos;

5. Representação legível por máquina das declarações centrais de uma contribuição.

Os benefícios das quatro primeiras exigências aqui mencionadas são, em grande parte, indiscutíveis. Predominantemente, existem também soluções técnicas que simplesmente precisam ser utilizadas e, para promover este uso, (melhor) apoiadas pelas infraestruturas de publicação existentes ou mais abertas para uso posterior por terceiros.

A codificação estruturada dos metadados relacionados a documentos (requisito 1) é atualmente tratada, de maneira predominante, separadamente dos próprios textos dos artigos nas bases de dados dos fornecedores, onde são, naturalmente, utilizados intensivamente para fins de descoberta. Uma melhor integração poderia ser implementada através da

incorporação dos metadados correspondentes na área "Propriedades" de um arquivo PDF. Outros métodos são manter o DOI como referência para a contribuição e os metadados correspondentes ou, se for utilizado um formato semiestruturado, codificar esses metadados na área correspondente do arquivo XML (por exemplo, em TEI ou JATS <jats.nlm.nih.gov/; e em TEI: tei-c.org>).

No caso da codificação explícita de estruturas de texto, por exemplo, atribuindo seções de texto a classes estruturais ou semânticas (requisito 2), ocorre que isto geralmente não é implementado no contexto de arquivos PDF (apesar das possibilidades realmente disponíveis). No formato PDF domina demais o aspecto do layout. As infraestruturas que poderiam utilizar tais informações não estão suficientemente desenvolvidas. Aqui, somos dependentes ainda das possibilidades de formatos baseados em XML como JATS ou TEI, que até agora desempenham só um papel marginal no setor de periódicos. Até o momento, apenas pouquíssimas revistas ou editoras aceitam manuscritos em tais formatos. As exceções a esta regra são o *Digital Humanities Quarterly* (DHQ) e o *Journal of the Text Encoding Initiative* (jTEI), que usam XML-TEI (Harrison, 2016). O fornecedor de periódicos *Public Library of Science* (PLOS) assim como *Open Library of Humanities* geram uma versão XML em JATS para seus artigos de periódicos e a oferecem para download. O LaTex é frequentemente aceito nas áreas das Ciências Naturais e da Informática. Elsevier, por exemplo, o converte internamente para XML, mas não o publica nesse formato.

Para a codificação estruturada das referências bibliográficas (requisito 3), por outro lado, existe toda uma gama de formatos de dados bem estabelecidos, entre os quais a BibTex certamente provou ser particularmente central. Numerosas ferramentas, tais como Citavi (acesso pago; <citavi.com> ou Zotero (gratuito; <zotero.org>) permitem o gerenciamento conveniente de tais dados, bem como sua utilização na redação de textos científicos. Hoje, porém, estes formatos e programas são usados principalmente para gerar uma bibliografia uniformemente formatada num estilo de citação específico. A bibliografia é então anexada ao texto, mas com perda da estruturação explícita dos dados. Além disso, conceitos existentes para a extensão de tais dados, por exemplo, por informações sobre o propósito de uma referência em uma publicação usando uma ontologia como a Citação Tipo Ontologia (CiTO; <zotero.org>), são pouco utilizados. Para fazer melhor uso desses dados, são necessárias adaptações infraestruturais de longo alcance, como a possibilidade de adicionar um arquivo BibTex com os dados bibliográficos a uma determinada publicação como suplemento. Também revistas, que aceitam a submissão de artigos em LaTex + BibTex geralmente publicam em PDF.

Dados Abertos Vinculados para codificação de conteúdo

Passemos agora aos dois últimos requisitos, que estão diretamente relacionados com o uso de Dados Abertos Vinculados. Desde o artigo de Shotton, a indústria editorial mudou drasticamente. No entanto, o que ele disse na época ainda se aplica: "com algumas exceções brilhantes, as revistas on-line atualmente não fornecem nenhuma margem semântica de texto que facilitaria uma maior compreensão do significado subjacente" (Shotton, 2009, p. 87). Os exemplos mencionados por Shotton não foram aceitos. A virada semântica da publicação científica ainda está por vir. Há certamente muitas razões para isto, entre elas provavelmente uma falta de consciência dos benefícios e das possibilidades de implementar tal codificação semântica. Este é o ponto do qual a presente contribuição gostaria de partir.

A leitura automática da indexação das entidades e conceitos de uma contribuição (requisito 4) não é algo inteiramente novo. Afinal, ela já é a base para a criação de um índice ou registro de palavras-chave, como é usual para livros de não ficção, geralmente com respeito a entidades (tais como pessoas, organizações, lugares e títulos de trabalho) por um lado, e a conceitos (resumos, conceitos, termos técnicos) por outro. O que é novo no contexto das publicações digitais, entretanto, é que a indexação não apenas conecta o registro dentro de uma publicação com as referências no texto e assim torna a publicação acessível, mas que as entidades e conceitos podem ser identificados de forma única, ligando-os a dados de controle de autoria (pt.wikipedia.org/wiki/Controle_de_autoridade) e integrados numa ontologia específica do domínio, e assim tornando-se parte da Web Semântica como Dados Vinculados (Abertos) (Linked [Open] Data) (Dengel, 2012). Além disso, tais marcações devem ser inseridas não apenas manualmente pelos autores, mas devem também ser usadas ferramentas disponíveis para anotação automática (por exemplo, através do reconhecimento da entidade nomeada) e para integração em ontologias relevantes.

Tal integração do conteúdo do artigo na Web Semântica não só torna possível a indexação através de numerosas publicações, mas as entidades e conceitos indexados também podem ser dinamicamente enriquecidos com informações adicionais: pessoas mencionadas, por exemplo, por dados de vida e local de ação ou disciplina(s). Tudo isso requer infraestruturas em um sentido múltiplo: no sentido de formatos de dados que permitam um enriquecimento correspondente das publicações; de conjuntos

de dados padrão que possam ser referenciados para a desambiguação e enriquecimento das entidades; e de infraestruturas de publicação que permitam a indexação, vinculação e utilização dos dados correspondentes. No que diz respeito aos formatos de dados, o JATS é limitado em sua expressividade, enquanto o TEI fornece todos os mecanismos essenciais. No que diz respeito à infraestrutura de publicação, o autor muitas vezes não tem conhecimento de nenhuma plataforma de publicação, editoras ou periódicos que aceitem os formatos de dados anotados correspondentes no momento da submissão e que, assim, também utilizariam os dados para publicação. No entanto, há muito a aprender de outras áreas de aplicação de dados normativos, como por exemplo, na edição filológica (STADLER, 2012; KAMZELAK, 2016). Por último, mas não menos importante, a integração na Web Semântica também requer acesso aberto para que o acesso livre às publicações relevantes possa ser feito em todos os locais de publicação e não limitado apenas ao portfólio de um fornecedor. Entretanto, isto está em conflito direto com os interesses dos editores que querem manter seus leitores em sua própria plataforma.

O último requisito mencionado acima é que uma publicação legível por máquina deve oferecer suas principais declarações ou resultados de forma cuidadosamente modelada semanticamente. Seringhaus e Gerstein (2007) chamam isto de "Resumo Digital Estruturado" (*Structured Digital Abstract*), o que definem como um "resumo XML legível por máquina de fatos pertinentes no artigo". Ao contrário das exigências discutidas até agora, esta exigência é geralmente menos aceita, pelo menos no contexto da redação de publicações científicas. Este estado de desenvolvimento está relacionado ao fato de que existem soluções técnicas menos específicas e suficientemente avançadas e que o tema em si tem sido conceitualmente muito menos bem refletido. Deve ser feita uma distinção entre a implementação técnica, por um lado, e a solução conceitual, por outro. A implementação técnica parece secundária em relação ao estado atual do debate. Ela é principalmente uma questão de construção de consenso e das ferramentas disponíveis em uma comunidade. Parece claro, entretanto, que tal implementação (como no caso de entidades e conceitos de rotulagem) deveria utilizar os mecanismos de *Linked Open Data* (LOD) e, portanto, da Web Semântica.

O enfoque, no que segue, será, portanto, no lado conceitual, provavelmente o aspecto mais controverso do tópico. No contexto literário, parte da dificuldade decorre também do seguinte: diferente da situação na biologia ou química, onde já foram desenvolvidas várias ontologias relevantes, ou da linguística, onde já existe uma vasta experiência e projetos

relevantes com a "Linguistic Linked Data Cloud" (linguistic-lod.org), na área do estudo de literatura, a utilização do Linked Open não está ainda muito enraizada – pelo menos não para além da codificação de metadados bibliográficos básicos e no desenvolvimento de edições digitais. Por esta razão, seguirão algumas reflexões nesta direção, utilizando um estudo de caso do campo da especialização do autor, a história literária.

O estudo de caso toma a perspectiva do enriquecimento retrospectivo de publicações científicas existentes por meio de *Structured Digital Abstracts*. A experiência a ser reunida aqui também será útil para responder à questão de como documentar o conteúdo de publicações científicas emergentes de uma forma legível por máquina. Em termos de conteúdo, o foco é o romance francês da segunda metade do século XVIII. Com base em uma bibliografia de todos os romances publicados na França entre 1750 e 1799 (existem cerca de 2.000 títulos diferentes) já modelada como um LOD (LÜSCHOW, 2020), as entidades (romances e romancistas) aí contidas foram enriquecidas com afirmações relevantes à história da literatura, assim sendo já acessíveis. As seguintes considerações fazem parte das considerações preliminares para o projeto *Mining and Modeling Text* (MiMoText), que está atualmente começando no Centro Trier de Humanidades Digitais da Universidade de Trier, coordenado pelo autor deste artigo e financiado no âmbito da iniciativa de pesquisa do estado da Renânia-Palatinado (kompetenzzentrum.uni-trier.de/de/projekte/projekte/m).

Um breve trecho de texto sobre o romance *Candide*, a partir de uma visão geral da história da literatura francesa de Erich Köhler, servirá de exemplo:

> *Candide* é a obra mais lida de Voltaire e provavelmente já o foi durante a vida do autor. Quando foi impresso pela primeira vez em Genebra em 1759, foi imediatamente proibido, mas somente com o resultado de que houve treze novas edições no mesmo ano. (KÖHLER, 1984, p. 8)

A ideia básica agora é registrar o conteúdo central de um tal texto na forma de proposições basais, formuladas em forma de declarações "sujeito-predicado-objeto" no sentido de Linked Open Data, registradas como os chamados triplos, por exemplo, em um formato como RDF ou Turtle (ver DENGEL, 2012).

Antes de tudo, uma codificação semântica e explícita das declarações em uma publicação científica pode ser baseada nas entidades e conceitos anotados cuja distinção no sentido do requisito 4 (ver acima) é pressuposta aqui. Portanto, com a abordagem de Linked Open Data, já existem mecanismos para se referir a entidades e conceitos e utilizá-los

como entidades em declarações. No estudo de caso aqui tratado, as entidades relevantes são então pessoas (concretamente: autores de romances ou de literatura técnica; aqui: Voltaire) bem como obras (especificamente: romances, ou artigos individuais, capítulos, ou publicações monográficas; aqui: *Candide*). O inventário das entidades concebíveis deve ser entendido como uma lista incompleta e, portanto, não deve ser codificado. Além disso, entidades conceituais também podem ser derivadas de campos literários básicos, tais como conteúdo, estilo, gênero, época etc. A anotação (em XML-TEI e usando identificadores de conjuntos de dados padrão como o VIAF ou o *Getty Thesaurus of Geographical Names*) poderia então ter a seguinte forma:

> <title type="work" ref="viaf:176620251">Candide</title> é a obra mais lida <persName type="author" ref="viaf:36925746"> de Voltaire</persName> e provavelmente já o foi durante a vida do autor. Quando <date>1759/date> foi lançado em <placeName type="city" ref="tgn:7007279">Genebra <placeName>, foi imediatamente banido, mas apenas com o resultado de ter sido reimpresso treze vezes no mesmo ano.

No entanto, existem trabalhos exploratórios menos fundamentados conceitualmente, que podem ser utilizados para formular entidades como essas. Naturalmente, esses trabalhos podem, em primeiro lugar (e trivialmente), simplesmente conectar o próprio texto com as entidades mencionadas no texto, tal como

> Köhler_1984 (viaf:174648806) HAS_SUBJECT Voltaire (viaf:36925746); Candide (viaf:176620251); Genebra (tgn:7007279)

Porém, com base nisso, a questão central é agora como o conteúdo do texto pode ser formalizado. Esta pergunta toca na compreensão fundamental de uma determinada disciplina, já que se trata de determinar que tipo de declaração uma comunidade científica de especialistas considera fundamental para um determinado domínio. O exemplo acima contém uma série de afirmações que seriam questionadas aqui e que são chamadas de afirmações pseudoformalizadas:

- Voltaire (viaf:36925746) IS_CREATOR_OF Candide (viaf:176620251)
- Candide (viaf:176620251) HAS_PUBLICATION_DATE 1759

- Candide (viaf:176620251) HAS_PUBLICATION_LOCATION Genebra (tgn:7007279)
- Candide (viaf:176620251) HAS_RECEPTION_INTENSITY alta
- Candide (viaf:176620251) HAS_RECEPTION_TIME imediato;a longo prazo
- Candide (viaf:176620251) HAS_LEGAL_STATUS censurado (1759)

Neste exemplo, os três primeiros itens não são muito mais do que metadados bibliográficos, tais como as que se encontram em catálogos ou bibliografias temáticas (e como já existem no nosso caso). Para alguns destes tipos de declarações, especialmente quando se trata de informação prosopográfica e bibliográfica, pode-se utilizar para a formalização ontologias existentes, por exemplo, Dublin Core (para creator, publisher, date, title, subject) ou as ontologias SPAR (para posterior modelação bibliográfica) (PERONI; SHOTTON, 2018; ver en.wikipedia.org/w/index.php?title=-Dublin_Core&oldid=922336659). No entanto, isto não vale para as declarações seguintes. A questão central é, portanto, como deve ser concebida uma ontologia de tipos de afirmação central para um determinado domínio científico (aqui: história literária como parte dos estudos literários) e como pode ser estabelecido um consenso sobre estes temas na comunidade dos pesquisadores. Quais informações de domínio (aqui: informações da história literária) deviam ser formuladas como declarações básicas?

Comparativamente indiscutível, semelhante às declarações bibliográficas básicas já mencionadas, deviam ser informação estabelecidas prosopográficas como a que se encontra na Wikidata, por exemplo:

- (pessoa) DATE_OF_BIRTH (data); HAS_DATE_OF_DEATH (data)
- (pessoa) OCCUPATION (profissão)
- (pessoa) RELIGION (religião)
- (pessoa) MOVEMENT (ideologia, visão de mundo, crença)

Declarações um pouco mais específicas do domínio, ainda oficialmente padronizadas dentro de uma ontologia, mas praticadas, por exemplo, na Wikidata, são as seguintes (wikidata.org/wiki/Wikidata:List_of_properties/work#Wikidata_property_related_to_works_of_fiction):

- (pessoa) INFLUENCED_BY (pessoa)
- (pessoa) AWARD_RECEIVED (prémio)
- (obra) GENRE (gênero)
- (obra) CARACTERES (nomes de personagens)
- (obra) NARRATIVE_LOCATION (Localização geográfica)
- (obra) SET_IN_PERIOD (período de tempo)
- (obra) DERIVATIVE_WORK (obra)
- (obra) INSPIRED_BY (obra)
- (obra) NARRATOR (nomes de personagens)

Aqui começa a ficar evidente que um exame sistemático deste tipo de declarações sob a forma de uma ontologia ainda está pendente. Por exemplo, "RELIGION" e "MOVEMENT" têm uma relação pouco clara com "INFLUENCED_BY" no nível da pessoa e com "INSPIRED_BY" no nível do trabalho. Para alguns aspectos, as taxonomias ou ontologias existentes poderiam ser reutilizadas, por exemplo, na área de títulos profissionais (históricos e atuais), por exemplo, conforme o sistema HISCO (*Historical International Standard Classification of Occupations*; ver VAN LEEUWEN; MAAS; MILES, 2004). Para outros aspectos, tais como gêneros literários, épocas, formas ou temas, não há recursos comparativamente formalizados e consensuais. Evidentemente, os predicados usados no Wikidata até hoje também não são de maneira alguma suficientes para uma descrição histórica adequada de obras literárias, autores e épocas. Somente com relação a obras literárias, por exemplo, as seguintes informações adicionais seriam relevantes:

- (obra) HAS_FORM (prosa|verso|outro)
- (obra) HAS_NARRATIVE_PERSPECTIVE (autodiegético|homodietico|heterodietico) – forma narrativa em obras narrativas
- (obra) HAS_DIALOGUE_PROPORTION (porcentagem) – Porcentagem de discurso direto em um trabalho narrativo, como uma porcentagem de palavras ou frases.
- (obra) HAS_STAGE_DIRECTIONS (porcentagem) – Proporção de direções de cena num trabalho dramático, em porcentagem de palavras

Evidentemente, esta lista está longe de ser conclusiva. Uma modelagem sistemática do domínio ainda está pendente. Entretanto, deve-se acrescentar, neste contexto, que as informações coletadas ou extraídas não são consideradas fatos, mas sim afirmações. Na medida em que cada afirmação é atribuída a uma fonte, ela representa a opinião de um especialista ou do estado da pesquisa no momento da publicação. Consequentemente, um sistema de informação que coleta um grande número de tais declarações também pode conter declarações contraditórias ou incompatíveis, sem que o inventário do sistema seja considerado inconsistente.

O projeto atual "Mining and Modeling Text" (MiMoText) trata apenas indiretamente da questão de como as publicações científicas devem ser acessadas no futuro. Entretanto, o objetivo central é identificar e modelar semanticamente um certo inventário de tipos de declaração histórica literária em um *corpus* de literatura especializada existente (especialmente mais antiga, mais concisa) seguindo a abordagem delineada e, em grande parte, automaticamente. A historiografia literária mais antiga também deve ser tornada visível novamente para ser utilizável em larga escala através da modelagem e publicação como Linked Open Data, usando ontologias específicas do domínio. Além disso, informação sobre romances raramente lidos deve ser pesquisada desta forma e incorporada ao sistema de informação da história literária resultante. Desde que uma quantidade suficientemente grande de literatura especializada indexada semanticamente esteja disponível, isto criará um sistema de informação bibliográfica-histórica que suporta uma gama de cenários de aplicação. Por exemplo, será possível determinar o histórico de recepção de uma determinada autora não só quantitativamente (por exemplo, através do número de publicações relevantes por ano), mas também para compreender o conteúdo da obra, analisando a evolução dos respectivos temas abordados, as tendências de avaliação ou os respectivos autores comparativos mobilizados. Também será possível, com base nas declarações contidas no sistema (em termos de conteúdo, estilo, avaliação, classificação etc.), identificar obras literárias que, de acordo com os critérios escolhidos em cada caso, tenham características comuns e, portanto, possam ser adequadas para uma análise posterior e comparativa.

Esta forma de indexação semântica bastante elaborada e retrospectiva tornar-se-ia supérflua no futuro se a literatura especializada recentemente publicada já publicasse suas declarações essenciais sob a forma de Linked Open Data. A identificação das entidades relevantes e a formulação das declarações correspondentes poderiam ser feitas pelos próprios autores, ou poderiam ser desenvolvidos procedimentos que o fizessem automaticamente com base no texto completo, o que poderia ser feito

utilizando métodos desenvolvidos há algum tempo no campo da mineração de argumentos e da anotação semântica automática. No futuro, uma maior precisão (com simultaneamente menos cobertura) poderia ser esperada se uma prática correspondente de escrever publicações científicas fosse estabelecida, semelhante à prática atual de especificar uma série de palavras-chave ou selecionar termos de uma ontologia científica ao submeter artigos ou capítulos. Entretanto, ainda há um longo caminho a percorrer até lá, e os futuros métodos de publicação científica e a indexação retrospectiva da história das áreas de pesquisa terão um papel nisso. Por um lado, uma prática contemporânea de anotação semântica de publicações científicas em combinação com o texto completo também produz dados de treinamento para o desenvolvimento de procedimentos automáticos. Por outro lado, a indexação retrospectiva também pode ajudar a especificar os requisitos para futuros procedimentos de anotação semântica de publicações científicas e para ontologias subjacentes, específicas do domínio.

De qualquer forma, a visão do autor é que, num futuro próximo, não formularemos mais resultados de pesquisa apenas em prosa, em linguagem natural. Não produziremos, distribuiremos e receberemos mais artigos ou livros como arquivos PDF. Esta prosa também não será sem conexão com a publicação associada de conjuntos de dados e o código de programação. Em vez disso, este texto em prosa será vinculado a códigos e conjuntos de dados relevantes, fornecido com metadados ricos, marcado em sua estrutura de texto, anotado com entidades e conceitos usando dados bibliográficos estruturados, e publicado na forma de declarações LOD. Não se pretende alegar que o texto contínuo formulado em prosa em linguagem natural se tornará obsoleto como resultado. Porém, no futuro, o texto contínuo não ficará mais sozinho – ele será incorporado em um contexto rico, legível por máquinas de dados, código, com metadados, dados de citação e declarações modeladas. (Um suplemento de dados para esta publicação foi arquivado em Zenodo.org sob o seguinte DOI: doi.org/10.5281/zenodo.3898418).

Referências

DENGEL, Andreas (org.). *Semantische Technologien: Grundlagen, Konzepte, Anwendungen*. Heidelberg: Spektrum, 2012.

DIGITAL HUMANITIES im deutschsprachigen Raum (org.). Digitales Publizieren. Digitales Publizieren / DHd-Arbeitsgruppe "Digitales Publizieren". Wolfenbüttel : Herzog August Bibliothek, 2016. Disponível em: <dhd-wp.hab.de/?q=content/working-paper-digitales-publizieren>. Acesso em: 4 jun. 2020.

Harrison, Melissa. Collecting xml at article submission at eLife: two steps forward, one step back? *In*: Journal Article Tag Suite Conference (JATS-Con). *Proceedings 2016* [Internet]. Bethesda (md): National Center for Biotechnology Information (us), 2016. Disponível em: <ncbi.nlm.nih.gov/books/NBK350147>. Acesso em: 5 jun. 2020.

Kamzelak, Roland S. Digitale Editionen im Semantic Web: Chancen und Grenzen von Normdaten, frbr und rdf. *In*: Richts, Kristina; Stadler, Peter (org.). *„Ei, dem alten Herrn zoll' Ich Achtung gern'"*: Festschrift für Joachim Veit zum 60. Geburtstag. München: Allitera, 2016, p. 423-435. Disponível em: <dx.doi.org/10.25366/2018.29>. Acesso em: 5 jun. 2020.

Köhler, Erich. *Vorlesungen zur Geschichte der Französischen Literatur: Aufklärung II*, org. rieger, Dietmar. Stuttgart: Kohlhammer, 1984.

Lippi, Marco; TORRONI, Paolo. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology* v. 16, n. 2, p. 1-25, 2016. Disponível em: <doi.org/10.1145/2850417>. Acesso em: 3 jun. 2020.

Lüschow, Andreas. *Bibliographie du genre romanesque français 1751-1800 – rdf Model*. Zenodo 2019. Disponível em: <doi.org/10.5281/zenodo.3401428>. Acesso em: 3 jun. 2020.

_____. Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane. *Jahrestagung des DHd-Verbands 2020: Spielräume*. Paderborn, 2020. Disponível em: <doi.org/10.5281/zenodo.3666689>. Acesso em: 3 jun. 2020.

Peroni, Silvio; Shotton, David. The SPAR ontologies. *In*: Vrandecic, D. *et al.* (org.) *The Semantic Web – ISWC 2018. ISWC 2018. Lecture Notes in Computer Science*, vol. 11137, 2018. Cham: Springer. Disponível em: <doi.org/10.1007/978>. Acesso em: 3 jun. 2020.

Piron, Florence. Qui sait ? Le libre accès en Afrique et en Haïti. 2020. Disponível em: <laviedesidees.fr/Qui-sait.html>. Acesso em: 3 jun. 2020.

Pooley, Jeff. The library solution: How academic libraries could end the APC scourge, 2020. Disponível em: <items.ssrc.org/parameters/the-library-solution-how-academic-libraries-could-end-the-apc-scourge>. Acesso em: 4 maio 2020.

Schöch, Christof. Open Access für die Maschinen. *In*: Effinger, Maria; Kohle, Hubertus (org.). *Die Zukunft des kunsthistorischen Publizierens. In*: arthistoricum.net, 2020. doi: <doi.org/10.11588/arthistoricum.663.c9210>. Acesso em: 4 maio 2020.

Seringhaus, Michael R.; Gerstein, Mark B. Publishing perishing? Towards tomorrow's information architecture. *bmc Bioinformatics* v. 8, n. 17, 2007. Disponível em: <doi.org/doi:10.1186/1471-2105-8-17>. Acesso em: 3 jun. 2020.

Shotton, David. Semantic publishing: The coming revolution in scientific journal publishing. *Learned Publishing* v. 22, p. 85-94, 2009. Disponível em: <doi.org/10.1087/2009202>. Acesso em: 3 jun. 2020.

Speicher, Lara; Armando, Lorenzo; Bargheer, Margo; Eve, Martin Paul; Fund, Sven; Leão, Delfim; Mosterd, Max; Pinter, Frances; Souyioultzoglou, Irakleitos. *operas* Open Access Business Models White Paper, 2018. Disponível em: <doi.org/10.5281/zenodo.1323707>. Acesso em: 3 jun. 2020.

Stadler, Peter. Normdateien in der Edition. *Editio: Internationales Jahrbuch für Editionswissenschaft*, v. 26, n. 1, 2012. Disponível em: <doi.org/10.1515/editio-2012-0013>. Acesso em: 5 jun. 2020.

Uren, Victoria; Cimiano, Philipp; Iria, José; Handschuh, Siegfried; Vargas-Vera, Maria; Motta, Enrico; Ciravegna, Fabio. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web* v. 4, p. 14-28, 2005.

Van Leeuwen, Marco h. d., Maas, Ineke, Miles, Andrew. Creating a Historical International Standard Classification of Occupations: An exercise in multinational interdisciplinary cooperation. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* v. 37, n. 4, p. 186-97, 2004. Disponível em: <doi.org/10.3200/HMTS.37.4.186-197>. Acesso em: 5 jun. 2020.

Wilkinson, Mark D.; Dumontier, Michel; Mons, Barend. The fair Guiding principles for scientific data management and stewardship. *Scientific Data* v. 3, n. 160018, 2016. Disponível em: <nature.com/articles/sdata201618>. Acesso em: 5 jun. 2020.