CSBCSet: Um conjunto de dados para uma década de CSBC, seus eventos e publicações

Silas Lima Filho[1], **Luiz Paulo Carvalho**[1], **José Antonio Suzano** [1], **Michele A. Brandão** [2], **Jonice Oliveira**[1], **Flávia Maria Santoro**[3]

[1]UFRJ – Rio de Janeiro, RJ – Brasil

[2]IFMG/UFMG – Belo Horizonte, MG – Brasil

[3]UERJ – Rio de Janeiro, RJ – Brasil

luiz.paulo.carvalho@ppgi.ufrj.br, silaslfilho@ppgi.ufrj.br

jose.suzano@matematica.ufrj.br, michele.brandao@ifmg.edu.br

jonice@dcc.ufrj.br, flavia@ime.uerj.br

Abstract. *In this paper, we present a dataset about a decade of publications in the Brazilian Computing Society Congress (CSBC), between 2013 and 2022. We specify the extraction, processing and adjusments, storage and opening of data, with its limitations, challenges and lessons learned . We analyzed the data and metadata scenario, perceiving positive and negative aspects of the scope. Finally, we forward proposals for potential applications of this dataset, and related threats to validity.*

Resumo. *Neste trabalho apresentamos um conjunto de dados acerca de uma década de publicações que compõe o Congresso da Sociedade Brasileira de Computação (CSBC), entre 2013 e 2022. Especificamos a extração, tratamento, armazenamento e abertura dos dados, com suas limitações, desafios e aprendizados. Analisamos o cenário sobre os dados e metadados, percebendo aspectos positivos e negativos do escopo. Finalmente, encaminhamos propostas de potenciais aplicações do conjunto de dados, e ameaças à validade relacionadas.*

1. Introdução

O CSBC é o maior evento dedicado à computação da América Latina, congregando pesquisadores e instituições de todo Brasil a fazer-pensar a computação, brasileira ou internacional, de maneira participativa e colaborativa. Um dos cernes do CSBC são seus eventos acadêmico-científicos, com dinâmicas tradicionais da cultura acadêmico-científica, e.g., chamadas de trabalhos, publicações, sessões de apresentação oral. Conforme pesquisadores publicam e comunicam seus trabalhos, a ciência computacional brasileira avança.

Realizados pela Sociedade Brasileira de Computação (SBC) desde 1980, os eventos base ou satélite do CSBC trazem mais do que apenas as últimas tendências, tópicos e interesses de pesquisa em computação brasileiras; eles agenciam uma rede de pesquisadores, que estão filiados a instituições de diversos tipos; avançando áreas de pesquisa agrupadas em eventos, através de pesquisas, materializadas em suas publicações, sumarizadas por resumos, com palavras-chave. Além dos tradicionais artigos científicos, dados

abundantes expõem o panorama da ciência computacional brasileira, associados às práticas de pesquisadores comunicadas no maior evento brasileiro desta área.

Neste trabalho elaboramos o *CSBCSet*, um conjunto de dados processados e tratados contendo integralmente os dados sobre uma década de publicações acadêmico-científicas formais nos eventos mais longevos realizados e sediados no CSBC até 2022, base ou satélite, no intervalo de 2013 até 2022. Extraímos os dados e metadados do repositório oficial da SBC, complementando os dados e metadados faltantes manualmente. Com estes dados e metadados podemos estruturar um panorama, centrado nas publicações, de uma década de CSBC.

Uma das qualidades centrais está no caráter censitário e completo dos dados, com processamento e tratamento minucioso, para encaminhar o melhor, em quesito de qualidade de dados, conjunto de dados possível sobre a realidade da década de 2013 até 2022 do CSBC. Apresentamos uma intenção meta-científica primária à representação deste cenário pelos dados; e experiências, desafios e lições através do método.

A Seção 2 apresenta trabalhos relacionados, direta ou indiretamente; a Seção 3 apresenta o método e seu protocolo, os aspectos éticos e a instância do método; a Seção 4 apresenta limitações, discussões e encaminhamentos, e considerações finais.

2. Trabalhos relacionados

Diversos trabalhos tratam do CSBC sob um viés meta-científico [Ioannidis 2018], com ênfases diversas. Um diferencial deste trabalho, e do *CSBCSet*, é sua abrangência de uma década, através dos dez eventos mais longevos do CSBC, com processamento e tratamento minuciosos. Aqui nos interessa especificamente as publicações que lidem com o CSBC, seus eventos e publicações. [Santana e Braga 2020] fazem uma análise *cienciométrica* da participação de mulheres no CSBC entre 2017 e 2019, sem disponibilizar a base de dados utilizada. [Lobato et al. 2021] analisam dez anos de BraSNAM através de abordagens de Análise de Redes Sociais; [Digiampietri et al. 2017] conduz trabalho similar, analisando os primeiros cinco anos do evento.

Utilizamos a abordagem de [Lobato et al. 2021] [1] para disponibilizar o *CSBCSet*, através do *Zenodo*. [Digiampietri et al. 2017] não disponibiliza a base de dados utilizada e apresenta autorias anônimas, diferente deste presente trabalho, no qual identificamos os nomes das pessoas autoras e justificamos esta escolha na Seção 3.1.

3. Método e protocolo

Como expõe a Figura 1, utilizamos diversos sistemas neste trabalho [2] : (i) *Web Scraper* e *Octoparse*, versões gratuitas, utilizados para extração de dados baseados em estruturas estáticas ou dinâmicas, de forma automatizada e com parametrização manual; *Google Sheets*, gratuito, com o uso de planilhas para estruturação, armazenamento e compartilhamento de dados, com funcionalidades de manipulação e tratamento, e.g., filtragem e estatística descritiva; e para trabalho colaborativo simultâneo. Após os ajustes finais, armazenamos a versão final no *Zenodo*.

[1] https://zenodo.org/record/5038638 [acesso 04-04-2023]

[2] https://webscraper.io/ - https://www.octoparse.com/ [acesso 04-04-2023]

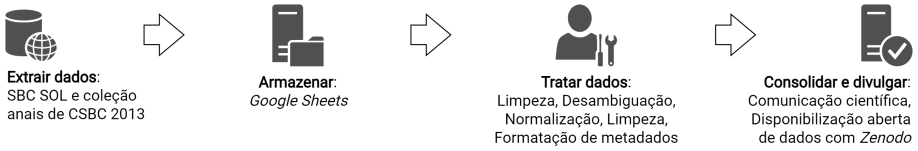

Figura 1. Principais etapas do método de realização deste trabalho.

Feita esta consideração, segue a análise racional para inclusão e exclusão de eventos. Coletamos os metadados de todas as publicações disponíveis entre 2013 – 2022 do SEMISH (Seminário Integrado de Software e Hardware); BraSNAM (*Brazilian Workshop on Social Network Analysis and Mining*); WIT (*Women in Information Technology*); WEI (Workshop sobre Educação em Computação); BreSCI (*Brazilian e-Science Workshop*); WCAMA (Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais); WPerformance (WPerf. – Workshop em Desempenho de Sistemas Computacionais e de Comunicação); SBCUP (Simpósio Brasileiro de Computação Ubíqua e Pervasiva); CTIC (Concurso de Trabalhos de Iniciação Científica); e CTD (Concurso de Teses e Dissertações). No total, extraímos dados e metadados de 97 edições de eventos, pois o WIT iniciou a chamada de trabalhos em 2016.

Excluímos os eventos que, em 2023, apresentam nove edições ou menos; os eventos com mais de dez edições entre 2013 – 2022 que desacoplaram do CSBC, e.g., WIM (Worskhop de Informática Médica); eventos que fomos incapazes de coletar todos os metadados de todos os trabalhos entre 2013 – 2022, e.g., WCGE (Workshop de Computação Aplicada em Governo Eletrônico) em sua décima edição em 2022; eventos com dinâmica epistêmica-prática não tradicional, e.g., JAI (Jornada de Atualização em Informática).

No processamento e tratamento dos dados, diferencial deste trabalho, limpamos a base de dados inicial, e.g., retirando espaços em excesso ou dados extraídos sem valor para intenção principal ou valores errôneos; realizamos a desambiguação dos nomes de pessoas autoras e suas filiações/instituições; definimos um padrão para os dados e formatamos os metadados; deliberamos sobre situações ambíguas ou conflitantes entre os dados; verificamos e revisamos a qualidade do conteúdo. Finalizada esta etapa, armazenamos e disponibilizamos o *CSBCSet* no formato CSV (*Comma Separated Values*), ODS (*OpenDocument Spreadsheet*), para usuários menos habituados à CSV, e JSON (*JavaScript Object Notation*).

3.1. Aspectos éticos

Extraímos esta discussão de [Carvalho et al. 2023], pela semelhança e aplicação do *CSBCSet*. Manter os nomes das pessoas autoras ou seguir com anonimato foi um dilema ético que nos acometeu. **Moralmente**, é ausente uma justificativa, fundamentação ou base que objetivamente **determine** um anonimato, pelas diretrizes institucionais governamentais que regem a ética em pesquisa brasileira [Brasil 2016] ou questões éticas gerais de pesquisa [ANPEd 2019, Bos 2020]. Pois lidamos com dados abertos; dados de promoção e publicidade de atividade-fim; sem qualquer dado pessoal sensível ou potencialmente prejudicial ou consequencialmente negativo às partes; consiste de uma pesquisa secundária que trata especificamente de dados meta-científicos; por fim, é ausente de qualquer juízo moral ou de valor primário associado aos dados.

Eticamente debatemos uma série de possíveis consequências negativas ou preju-

diciais materiais e concretas, além da ética relativista (e.g., "não quero" ou "não gosto"), e fomos incapazes de pontuar justificativas morais significativas o bastante para seguir com o anonimato. Adicionalmente, pessoas autoras podem utilizar destes dados ou informações para sua vantagem, e.g., em memoriais de promoção de carreira.

3.2. A construção do conjunto de dados, o método na prática

Iniciando pela **extração**, extraímos dados da biblioteca digital SBC SOL, repositório oficial da SBC. A extração foi feita no mês de fevereiro de 2023. A estrutura sucinta, objetiva e bem estruturada da SBC SOL permitiu uma extração rápida e simples. Os dados das edições do WCAMA, WPerf. e WEI de 2013 estavam indisponíveis, e foram extraídos manualmente da coleção de anais do CSBC 2013. O motivo central de estabelecermos este intervalo de tempo, 2013 – 2022, é pela quantidade e qualidade dos dados disponíveis na SBC SOL. As publicações de 2012 e abaixo estão, em sua maioria, em coleções de arquivos com formatos inadequados e extensos, de difícil extração, com caracteres codificados com problemas e sem uma estrutura sintática padrão uniforme, dentre outros problemas. Os dados extraídos em CSV foram **armazenados**.

Tabela 1. Detalhamento dos dados e metadados estruturados

Nome do campo	Formato	Exemplo	Descrição
Ano	Inteiro	2019	Ano da publicação
Evento	String	WIT	Nome do evento no qual o artigo foi publicado
Edição	Inteiro	13	Quantitativo da edição na qual o artigo foi publicado
Título	String	"Gêneros e suas nuances no ENEM"	Título completo do artigo
Pessoa autora	String	"Cristina Ciferri"	Nome da(s) pessoa(s) autora(s) do artigo
Sexo	String	"F"	Sexo da pessoa autora [F XOR M]
Instituição 1	String	"USP"	Primeira filiação da pessoa autora
UF Inst. 1	String	"SP"	Unidade Federativa da 1ª filiação institucional da pessoa autora [LISTA COM 27 UF]
Instituição 2	String	"-"	Segunda filiação da pessoa autora
UF Inst. 2	String	"-"	Unidade Federativa da 2ª filiação institucional da pessoa autora (caso haja) [LISTA COM 27 UF]
Idioma	String	"pt-br"	Idioma da publicação [en XOR pt-br XOR esp]
Resumo	String	"Visando investigar a diferença de gêneros [...]" (trecho do original)	Resumo do artigo
Palavra-chave	String	"Exame Nacional do Ensino Médio, ENEM, gênero, ciências exatas, desempenho dos participantes"	Palavras-chave do artigo

Logo após segue o **tratamento** dos dados. Como a estrutura permitia e o enquadramento era limitado, extraímos todas as categorias de dados possíveis das páginas, e.g., data de publicação ou link de acesso. Na **limpeza**, nós excluímos dados e metadados fora do escopo intencionado, e.g., palavras-chave do BreSci 2015 contendo apenas o caracter "a". A Tabela 1 expõe o resultado final da **formatação dos metadados**. Apenas uma publicação do SEMISH 2015 foi excluída, sem dados de autoria ou filiações, "Desmassificando a Educação utilizando IOT para Construir Games Inteligentes Personalizados" (ausente na página de informações e no arquivo da publicação).

O idioma das publicações foi inferido a partir dos resumos, caso ausente o resumo, do título. Mesmo ausentes textos em espanhol no conjunto de dados, possibilidade futura desta ocorrência é uma realidade. Utilizamos a mesma abordagem de [Lobato et al. 2021] para associar o sexo [3] às pessoas autoras. Utilizamos uma base de dados online [4] para

[3]Reconhecemos a distinção entre sexo e gênero, mas o sistema tende a associar esses rótulos. Isso é relevante para a pesquisa sobre a participação feminina na computação brasileira [Santana e Braga 2020].

[4]https://gist.github.com/alexandremcosta/c9361cc23722a5aa1133 [acesso 04-04-2023]

associar as UFs (Unidades Federativas) às instituições. Idioma, sexo, e UF da instituição são dados secundários, enriquecendo a proposta e complementando os dados originários. Campos de dados não informados, ausentes ou desconhecidos são preenchidos com "-".

O problema ocorreu com pessoas autoras e instituições, e seguimos para a **normalização** e **desambiguação**. Iniciando pelas pessoas autoras, além do problema "mesma pessoa, sintaxe diferente" [Digiampietri et al. 2015], encontramos nomes categoricamente errados, e.g., Valderi Leithardt encontra-se na página da publicação como "Leithhardt", o que em um cenário de análise computacional seriam consideradas duas pessoas diferentes. O problema tradicional de ambiguidade de nomes se torna mais complexo em português brasileiro devido à acentuação, então Aletéia Araújo apresenta diversas variações, e.g., com as duas acentuações, acentuação em apenas uma letra, ou simplesmente sem acentuação alguma. Os problemas de erros resolvemos rapidamente, presentes em sua maioria em nomes estrangeiros atípicos na cultura brasileira; para ambiguidade, selecionamos o nome mais simples e com maior ocorrência entre todos, e desambiguamos com base nele. Nos precavemos para evitar que esta simplificação gerasse novas ambiguidades. Em último caso, recorremos à publicação para resolver conflitos.

Nas instituições há um problema semelhante a pessoas autoras. Pessoas autoras em instituições às quais nunca estiveram vinculadas/filiadas e casos de discrepância entre página de informações da SBC SOL e a publicação, até casos onde ambos estão errados. Normalizamos os nomes das instituições, utilizando os padrões nos *websites* oficiais das mesmas, e.g., "puc minas" para "PUC Minas". Em caso de conflito de instituições resolvemos da seguinte forma: intervalo de um ano de diferença com instituições diferentes e mesma pessoa, ignora, e.g., alteração de filiação entre 2015 e 2017; valor imediato, buscar no currículos Lattes para correção, em caso de conflito, manter valor da filiação anterior; mesmo ano, mais de uma publicação e filiações diferentes entre elas, novamente busca no currículo Lattes, em caso de conflito, mantém a opção mais simples, com menos filiações.

Há inconsistência na ocorrência de palavras-chave entre eventos, edições e quantidade de publicações. Algumas edições de eventos têm publicações com palavras-chave, enquanto outras não. Apenas no BraSNAM todas as publicações de 2013 têm palavras-chave, mas em outros eventos isso ocorre em apenas um ou poucos casos. A maioria das publicações só apresenta palavras-chave de forma ampla e, às vezes, incompleta, a partir das edições de 2020.

4. Limitação, discussão e considerações finais

Como etapa final, **disponibilizamos** online e abertamente o *CSBCSet* [5], [acesso 27-05-2023] , respeitando os princípios de ciência aberta. Traz 7559 registros, representando 1997 publicações, 4961 pessoas autoras, 415 instituições. O *CSBCSet* está limitado aos dez eventos extraídos e no respectivo intervalo temporal, 2013 - 2022. Mesmo que deixe de representar o CSBC em sua totalidade nestes anos, traz contribuições de dados processados e tratados para futuras aplicações. Trabalhos futuros podem incluir a expansão do *CSBCSet* para mais eventos, anos ou outros tipos de comunicação informais. Quanto à validade dos dados, confiamos em sua veracidade e os processamos para melhorar sua qualidade e compatibilidade com a realidade. Após extensas revisões, acreditamos que há uma probabilidade de erros ou problemas em menos de 1% dos registros (cerca de 75).

[5]https://zenodo.org/record/7977462

Recomendamos revisões manuais após abordagens automatizadas. Os dados contêm informações valiosas sobre o trabalho das pessoas autoras, relevantes para sua prática especializada e profissional. A desambiguação de autorias e instituições é sensível, e automatizá-la traz riscos significativos, como mudança de filiação institucional ou ambiguidade em nomes comuns, como João da Silva. Esses dados podem ser usados para análise posterior pela SBC, incluindo normalização e tratamento dos dados.

Tendo superado os problemas, desafios e limitações, apresentamos o *CSBCSet*, com dados de publicações formais de uma década dos eventos mais longevos do CSBC, habilitando diversas pesquisas posteriores, com dados tratados minuciosamente.

5. Agradecimentos

O presente trabalho foi apoiado pela CAPES – Edital nº 09/2020 - Proc. nº 223038.014313/2020-19, e parcialmente apoiado pelo programa *Oracle for Research* (nº prêmio CPQ-2160239).

Referências

ANPEd (2019). Ética e pesquisa em educação: subsídios – volume 1. volume 1. ANPEd, Rio de Janeiro, RJ.

Bos, J. (2020). *Research Ethics for Students in the Social Sciences*. Springer Cham, 1st edition.

Brasil (2016). Ministério da saúde. RESOLUÇÃO Nº 510, DE 07 DE ABRIL DE 2016. Disponível em: `https://4658.short.gy/CEP_2016` [acesso 27/02/2023].

Carvalho, L. P., Lima Filho, S., Suzano, J., Brandão, M., Oliveira, J., e Santoro, F. M. (2023). Uma década de interações entre eventos e pesquisadores do csbc: Um estudo meta-científico. In *Anais do XII Brazilian Workshop on Social Network Analysis and Mining*, Porto Alegre, RS, Brasil.

Digiampietri, L., Linden, R., e Barbosa, L. (2015). Desambiguação de nomes em redes sociais acadêmicas: Um estudo de caso usando DBLP. In *Anais do IV BraSNAM*, Porto Alegre, RS, Brasil. SBC.

Digiampietri, L., Mugnaini, R., Pérez-Alcázar, J., Delgado, K., Tuesta, E., e Mena-Chalco, J. (2017). Análise da evolução, impacto e formação de redes nos cinco anos do brasnam. In *Anais do VI BraSNAM*, Porto Alegre, RS, Brasil. SBC.

Ioannidis, J. P. A. (2018). Meta-research: Why research on research matters. *PLoS Biol*, 16(3).

Lobato, F., Sousa, G., e Jr., A. J. (2021). Brasnam em perspectiva: uma análise da sua trajetória até os 10 anos de existência. In *Anais do X BraSNAM*, pp. 217–228, Porto Alegre, RS, Brasil. SBC.

Santana, T. e Braga, A. (2020). Uma análise cienciométrica das publicações do congresso da sociedade brasileira de computação na perspectiva das mulheres na computação. In *Anais do XIV WIT*, pp. 279–283, Porto Alegre, RS, Brasil. SBC.