INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA

NOTA TÉCNICA Nº 5/2021/CGCQTI/DEED

PROCESSO Nº 23036.004686/2020-58

1. **ASSUNTO**

1.1. Informa os resultados e recomendações do Termo de Execução Descentralizada nº 8750 – Controle de Privacidade nos Censos Educacionais do Inep (TED 8750), e sugere encaminhamentos.

2. **SUMÁRIO EXECUTIVO**

2.1. Trata a presente Nota Técnica da apresentação dos resultados finais e recomendações do TED 8750, executado pela Universidade Federal de Minas Gerais (UFMG), em cumprimento ao plano de trabalho do Termo de Execução Descentralizada (TED 8750), processo SEI 23036.007050/2019-24.

2.2. Refere-se, ainda, à proposição de encaminhamentos para o tratamento das fragilidades apontadas nos microdados públicos dos Censos Educacionais, e provavelmente nos microdados públicos de outras pesquisas, avaliações e exames, divulgados no site do Inep, a serem observados pela gestão do Inep frente aos limites éticos, metodológicos e também legais trazidos pela Lei Geral de Proteção de Dados Pessoais – LGPD (Lei nº 13.709, de 14 de agosto de 2018).

3. **ANÁLISE**

3.1. O avanço das tecnologias de informação e comunicação (TIC), associado às mudanças econômicas e dos processos produtivos sustentados na digitalização das atividades humanas, bem como a institucionalização de novas leis e regulamentos para o tratamento de dados pessoais impõem desafios às diversas instituições de pesquisa em todo o mundo, visando assegurar os diferentes compromissos técnicos e legais requeridos na produção das estatísticas oficiais: desde a expectativa dos cidadãos de que a informação seja tratada exclusivamente para a finalidade a que se propõe, com o devido tratamento da privacidade, até as exigências que se impõem para que as estatísticas ofereçam informações oportunas, adequadas para o enfrentamento dos problemas contemporâneos das sociedades.

3.2. Nesse contexto, essas organizações são compelidas a demonstrar que os dados por elas divulgados não permitem a identificação ou reidentificação dos indivíduos que fazem parte do público-alvo das pesquisas, requerendo, por vezes, a atualização dos procedimentos técnicos e dos produtos de divulgação estatística. Ao mesmo tempo, esforçam-se para desenvolver novos produtos informacionais para atender demandas cada vez mais recorrentes da sociedade e dos governos por novas informações e prazos mais curtos de atualização.

3.3. A questão é crucial para o desenvolvimento das pesquisas sociais, entre elas compreendida a pesquisa estatística, para as quais a confidencialidade dos dados pessoais tratados é um pressuposto ético e um requisito metodológico, e no caso da pesquisa estatística é tratada objetivamente em Resolução Internacional da Assembleia Geral das Nações Unidas (Resolução nº 68/261, de 29 de janeiro de 2014), em manuais e documentos técnicos dos sistemas estatísticos nacionais e em normas legais nacionais, no caso brasileiro, entre outras: Lei nº 5.534/1968, Lei nº 5.878/1973, Lei de Acesso à Informação - LAI (Lei 12.527/2011) e Lei Geral de Proteção de Dados Pessoais - LGPD (Lei 13.709/2018).

3.4. O comprometimento de dados pessoais controlados por empresas e órgãos da Administração Pública e os riscos associados ao acesso e ao uso impróprio dessas informações é uma questão que tem ganhado cada vez mais a atenção da sociedade, culminando em vários países na elaboração de legislações gerais sobre o tratamento de dados pessoais tanto pelo setor privado quanto pelo setor público. Mesmo após a promulgação da LGPD e até em decorrência dela, os impactos seguem

requerendo novas abordagens, ainda no campo jurídico, como o processo legislativo de Proposta de Emenda à Constituição 17/2019. No âmbito da atuação técnica do Inep, por sua vez, o esforço é assegurar os compromissos assumidos de modo a sustentar a confiança e a cooperação dos indivíduos, fatores necessários ao desenvolvimento das pesquisas estatísticas, para a validade dos seus resultados e, em ulterior análise, para a sustentação da capacidade técnica da autarquia cumprir suas atribuições legais.

3.5. Os riscos, entretanto, traduzem-se em fatos que expõem continuamente a criticidade da questão. Um mapa de ameaças cibernéticas, atualizado continuamente, elaborado por uma empresa de soluções de segurança da informação, apresenta a "indústria do setor educacional" como um alvo preferencial de ataques (para mais informações consultar https://threatmap.checkpoint.com). Não raro, depara-se com notícias na imprensa sobre o comprometimento de dados pessoais relacionados ao setor, dentre os quais destaca-se abaixo alguns casos ilustrativos recentes:

- Falha na Secretaria da Educação do DF expõe dados de quase 1,5 milhão de alunos.
- Falha em sistema do MEC expõe dados até de Bolsonaro e Lula.
- Vazam emails, senhas e fotos de milhares de estudantes do Pernambuco.
- Servidor desprotegido da UEESP expõe dados de milhares de universitários paulistas.
- Exclusivo: falha em sistema de franquia de escolas expõe dados de alunos, funcionários e franqueados.
- Dados sigilosos de 3,8 milhões de estudantes de SP vazam na internet.

> *"...Um incidente, particularmente se recebe forte atenção da mídia, tem um impacto significante sobre a cooperação dos respondentes e consequentemente sobre a qualidade das estatísticas oficiais" (Dupriez, O.; Boyko, E. International Household Survey Network – IHSN, 2010. Tradução. Livre).*

3.6. Embora o risco de identificação de pessoas nos microdados dos censos da educação já fosse conhecido, desde 2016, a partir do trabalho acadêmico de Maria Jane de Queiroz e Gustavo Motta, publicado publicado em 2015 (Privacidade e Transparência no Setor Público: Um Estudo de Caso da Publicação de Microdados do INEP), a soluções apresentadas no estudo não solucionam de maneira definitiva o problema. Desde então, a equipe técnica da Diretoria de Estatísticas Educacionais tem se debruçado mais detidamente sobre a questão, o que tem requerido, para além do desenvolvimento de habilidades técnicas, uma discussão mais aprofundada sobre questões éticas, históricas e metodológicas da pesquisa estatística, assim como acerca do debate recente sobre privacidade e tratamento de dados pessoais, carreado nas discussões públicas do processo legislativo de formulação da Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/2018). Com o tempo, as evidências do problema passaram a fatos eventuais demonstrando a fragilidade anteriormente identificada, como se pode verificar, por exemplo, nas matérias: Estudo inédito indica alta chance de fraude em mil provas do Enem e Dados de Filipe Sabará não aparecem em censo do MEC. A mensuração objetiva desse risco, entretanto, não era do alcance da equipe técnica do Instituto.

3.7. Considerando o contexto apresentado, a necessidade de adequação do tratamento de dados pessoais aos requisitos trazidos pela LGPD, e a necessidade de manutenção da relação de confiança entre o INEP e o público-alvo de suas pesquisas, especialmente os estudantes, suas famílias e toda a cadeia de informantes e usuários das estatísticas produzidas; esta Diretoria observou a oportunidade e a necessidade de se avaliar e, eventualmente, atualizar os métodos de controle de divulgação estatística utilizados na divulgação dos resultados de suas pesquisas, os Censos da Educação Básica e da Educação Superior. Nessa esteira, enfim, foi elaborado e executado o projeto do TED 8750, a partir da identificação de um parceiro com conhecimento científico reconhecido e capacidade técnica suficiente para a realização de procedimentos experimentais que mensurassem objetivamente os riscos de identificação dos titulares de dados pessoais tratados pelas pesquisas e disseminados nos microdados públicos, além da realização de estudos de caso de outras organizações com atribuições semelhantes às do Inep, e da aplicação de técnicas referenciadas na literatura científica, os quais pudessem orientar a atuação do Instituto.

3.8. Os estudos das referências bibliográficas apontaram que nas últimas décadas a comunidade científica tem se dedicado intensamente à pesquisa na área de controle de divulgação estatística, com técnicas cada vez mais refinadas sendo elaboradas. Como exemplos, no campo da anonimização, as técnicas determinísticas prevalentes incluem *k-anonymity*, *l-diversity* e *t-closeness*, enquanto no campo de sanitização probabilística de dados, as técnicas predominantes se baseiam em *differential privacy* e suas variantes (como *local differential-privacy*). Em tempo, a literatura assume que, até o momento, não há uma única técnica que se aplique idealmente a todos os cenários. Dada a complexidade e as nuances do assunto, o acompanhamento de uma equipe de *experts* é fundamental para a escolha da metodologia adequada com aplicabilidade adequada para o caso do Inep.

3.9. Os achados dos levantamentos e os resultados dos experimentos realizados no âmbito do TED 8750 encontram-se reunidos no documento Excertos de Resumos Executivos dos Produtos do TED 8750 (documento SEI nº 0696118), entre os quais destacam-se:

- *As atuais técnicas de proteção de privacidade utilizadas pelo Inep nos microdados divulgados, consistindo apenas em desidentificação – em que se removem possíveis identificadores individuais óbvios dos registros, como nome, CPF e RG – e em pseudonimização – em que tais identificadores individuais óbvios são substituídos por um código único de identificação artificialmente criado – estão sujeitas a vários riscos de privacidade já identificados na literatura.*

- *No caso do Censo da Educação Superior de 2018:*
  - *uma combinação de 3 quaseidentificadores (dia e ano de nascimento, e código do curso) permite ao adversário reidentificar com absoluta certeza até 39% dos indivíduos na base, ou aproximadamente 4.200.000 estudantes;*
  - *uma combinação de 4 quaseidentificadores (dia, mês e ano de nascimento, e código do curso) pode levar à reidentificação de até 80% dos indivíduos, ou aproximadamente 8.600.000 estudantes;*
  - *na mesma base a mesma combinação de 3 quaseidentificadores permite ao adversário reidentificar um indivíduo aleatoriamente selecionado como alvo com probabilidade de até 56% após a realização do ataque; enquanto que a mesma combinação de 4 quaseidentificadores eleva essa probabilidade a até 87%.*

- *No caso do Censo da Educação Básica de 2019:*
  - *uma combinação de 3 quaseidentificadores (mês e ano de nascimento e código da escola em que estudo) permite ao adversário reidentificar um indivíduo aleatoriamente selecionado como alvo com probabilidade de até 29,64% após a realização do ataque;*
  - *enquanto uma combinação de 4 quaseidentificadores (mês e ano de nascimento, município de nasciento e código da escola em que estudo) pode elevar essa chance de sucesso a até 49,86%.*
  - *Já o uso de todos os 10 quaseidentificadores (mês e ano de nascimento, município de nascimento e endereço, nacionalidade e país de origem, sexo, cor/raça, código da escola e dependência administrativa da escola) eleva o risco a 75,51%.*

- *O risco de inferência de atributo é sensivelmente maior que o risco de reidentificação do indivíduo.*

- *Os resultados obtidos demonstram de forma inequívoca que a atual forma de divulgação dos Censos Educacionais pelo Inep submete os titulares dos dados a consideráveis riscos de violação de privacidade, incluindo reidentificação e inferência de atributos sensíveis. Tal situação poderia constituir em violação da Lei Geral de Proteção de Dados Pessoais.*

- *A sanitização por amostragem reduziu significativamente a degradação determinística de privacidade, mas teve pouco efeito na degradação probabilística. O resultado está em conformidade com o conhecimento da literatura técnica: a amostragem diminui o número absoluto de indivíduos expostos a risco, entretanto aqueles presentes na amostra ainda*

podem estar sujeitos a graves violações. Por outro lado, a amostragem demonstrou uma manutenção razoável da utilidade estatística.

- *A sanitização por privacidade diferencial local praticamente eliminou a degradação probabilística de privacidade. (a ela não se aplicam análises determinísticas). Por outro lado, o método em sua forma atual apresentou um custo elevado –e potencialmente proibitivo– na utilidade estatística dos dados divulgados. Os resultados confirmam o conhecimento da literatura técnica: na maioria dos casos um trabalho longo e laborioso de ajuste de parâmetros e técnicas normalmente é necessário para se atingir o potencial máximo de utilidade da privacidade diferencial local.*

3.10. Os relatórios completos de cada um dos produtos desenvolvidos nos âmbito do TED 8750 encontram-se disponíveis no processo 23036.007050/2019-24, nos quais o tratamento de cada objeto é tratado de forma detalhada e exaustiva até possibilitar as conclusões exaradas. Frente aos resultados alcançados, as recomendações da equipe da UFMG para o Inep centram-se em duas ações:

- *abandono "temporário" da divulgação pública de microdados individualizados – exceto em salas seguras em favor da divulgação de dados agregados sanitizados por privacidade diferencial; e*

- *desenvolvimento de um método de sanitização de microdados por privacidade diferencial local desenhado e refinado especificamente para Censos Educacionais. Opção que demandaria tempo e recursos significativamente maiores, incluindo atividades de pesquisa e desenvolvimento.*

3.11. A equipe técnica da autarquia, em conjunto com a equipe de execução do projeto, concluiu que os estudos foram fundamentais para: (1) retornar ao Inep informação crucial para tomadas de decisões futuras, balanceando soluções de curto prazo – implementáveis de forma relativamente rápida, mas potencialmente sub-ótimas quanto à utilidade provida – e soluções de longo prazo – mais úteis, porém que demandem maior investimento de tempo e pessoal especializado; e (2) subsidiar uma comunicação efetiva com as partes interessadas sobre os desafios técnicos na divulgação de microdados, facilitando a compreensão sobre eventuais decisões futuras.

4. DOCUMENTOS RELACIONADOS

4.1. TED 8750 - PRICE Privacidade nos Censos Educacionais. Processo 23036.007050/2019-24: Produto 1 (SEI 0548943), Produto 2 (SEI 0571714), Produto 3 (SEI 0596198), Produto 4 (SEI 0653551), Produto 5 (SEI 0653553), Produto 6 (SEI 0653559), Produto 7 (SEI 0676339), e Produto 8 (SEI 0678060).

4.2. As demais referências externas citadas apresentam, todas, links para acesso ao conteúdo completo.

5. CONCLUSÃO

5.1. Nessa perspectiva, a Diretoria de Estatísticas Educacionais ressalta que os riscos já conhecidos, os fatos recentes e os resultados dos estudos realizados apontam que a forma adotada atualmente pelo Inep para o tratamento da privacidade nos microdados públicos é ineficaz. Ante essa constatação, sugere-se um conjunto de ações para remediar o problema com ações de curto, médio e longo prazos, tendo por perspectiva a persecução de soluções que possibilitem o retorno da divulgação de microdados públicos com segurança e aderente aos princípios éticos, metodológicos e legais vigentes.

5.2. Nesse sentido, propõe-se a relação das seguintes atividades:

- Solicitar análise jurídica à PROJUR se, frente ao conhecimento técnico produzido pelo TED, a sustentação da disseminação de microdados no seu formato atual, fere princípios legais vigentes.

- Apresentar o estudo realizado (TED) à ANPD e requerer apoio técnico para o desenvolvimento de ações intensivas e de curto prazo para o alinhamento das atividades do Inep aos requisitos da LGPD, incluindo a retirada do ar dos microdados de pesquisas,

avaliações e exames que contenham dados individualizados e pessoais, ainda que na forma pseudinimizada, uma vez que ficou comprovado que são dados identificáveis.

- Propor um novo TED à UFMG para dar seguimento às recomendações exaradas, em especial o tratamento de dados agregados por privacidade diferencial e o aperfeiçoamento da solução para divulgação de microdados públicos tratados por privacidade diferencial.

- Incluir a questão para o tratamento no âmbito do projeto estratégico da reestruturação da disseminação do Inep (Plano Estratégico 2020-2023).

- Realizar seminário com os diferentes perfis de usuário para apresentar o estudo do TED, as necessidades de adequação da forma atual de disseminação do Inep à legislação vigente, as alternativas de mitigação do impacto sobre o uso dos microdados, bem como as ações que serão desenvolvidas para o reestabelecimento da divulgação dos microdados de forma a atender os requisitos legais e relação otimizada entre privacidade e transparência na comunicação científica.

- Reforçar e ampliar a capacidade de atendimento do Sedap.

- Ampliação da oferta de soluções de disseminação de dados de maneira dinâmica para reduzir o impacto da retirada dos microdados do ar para o perfil de usuários que precisam de gerar estatísticas diretas da pesquisa (requer o investimento em solução de visualização de dados, não apenas a aquisição, mas na sua implementação para processos e fluxos de comunicação externa das estatísticas, bem como a qualificação técnica dos servidores e demais colaboradores na sua utilização).

- Desenvolver estudos sobre a possibilidade de desenvolvimento de solução remota (não presencial) para o Sedap, ainda que com algumas limitações, para reduzir a pressão sobre o serviço presencial e como alternativa a usuários que poderiam ter suas necessidades atendidas pelo serviço.

- Incentivar e reforçar os esforços para o desenvolvimento da capacidade técnica do Instituto, por meio de realização de ações de capacitação e de pesquisa e desenvolvimento na área.

- Por em curso a implementação do extenso conjunto de ações coordenadas as serem desenvolvidas e implementadas para adequação aos requisitos legais da LGPD, necessários não apenas no âmbito da divulgação dos resultados.

5.3. Os pontos tratados nesta Nota Técnica, incluindo os resultados e recomendações do TED 8750, foram abordados em reunião do Comitê de Governança Institucional do Inep do dia 03 de maio de 2011, para os quais solicita-se o patrocínio e o comprometimento da alta administração com as ações apresentadas, uma vez que os impactos da questão têm repercussão direta sobre a capacidade técnica do instituto sustentar o cumprimento de suas competências específicas e correspondentes atribuições legais.

Fábio Pereira Bravin

Coordenador-Geral de Controle de Qualidade e Tratamento da Informação

De acordo,

Carlos Eduardo Moreno Sampaio

Diretor de Estatísticas Educacionais

Documento assinado eletronicamente por **Fábio Pereira Bravin**, **Coordenador(a) - Geral**, em 20/05/2021, às 18:14, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Carlos Eduardo Moreno Sampaio**, **Diretor(a)**, em 20/05/2021, às 19:48, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.inep.gov.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0697449** e o código CRC **D1469CFD**.

Referência: Processo nº 23036.004686/2020-58 SEI nº 0697449