



MINISTÉRIO DA JUSTIÇA E SEGURANÇA PÚBLICA
ARQUIVO NACIONAL

PLANO DE TRABALHO DO TERMO DE EXECUÇÃO DESCENTRALIZADA Nº 01/2022

1. DADOS CADASTRAIS DA UNIDADE DESCENTRALIZADORA

a) Unidade Descentralizadora e Responsável

Nome do órgão ou entidade descentralizador(a): **Arquivo Nacional**

Nome da autoridade competente: **Joelson da Silva**

Número do CPF: **013.000.987-32**

Nome da Coordenação Responsável pelo acompanhamento da execução do objeto do TED: **Coordenação-Geral de Tecnologia da Informação do Arquivo Nacional (COTIN)**

b) UG SIAFI

Número e Nome da Unidade Gestora - UG que descentralizará o crédito: **200247 - Arquivo Nacional**

Número e Nome da Unidade Gestora - UG Responsável pelo acompanhamento da execução do objeto do TED: **200247 - Arquivo Nacional**

2. DADOS CADASTRAIS DA UNIDADE DESCENTRALIZADA

a) Unidade Descentralizadora e Responsável

Nome do órgão ou entidade descentralizador(a): **Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)**

Nome da autoridade competente: **Cecília Leite Oliveira**

Número do CPF: **339.327.861-49**

Nome da Secretaria/Departamento/Unidade Responsável pelo acompanhamento da execução do objeto do TED: **Coordenação-Geral de Tecnologias de Informação e Informática (CGTI)**

b) UG SIAFI

Número e Nome da Unidade Gestora - UG que descentralizará o crédito: **240121**

Número e Nome da Unidade Gestora - UG Responsável pelo acompanhamento da execução do objeto do TED: **240121**

3. OBJETO

Definição de metodologia e aplicação de soluções informacionais e tecnológicas voltadas para o reconhecimento e transcrição paleográfica de documentos históricos no âmbito do Arquivo Nacional.

4. DESCRIÇÃO DAS AÇÕES E METAS A SEREM DESENVOLVIDAS NO ÂMBITO DO TED

Metas	Atividades	Indicador Físico	Prazo
1. Analisar e identificar bibliografia relacionada à Técnicas de Machine Learning	1.1 Habilitação da pesquisa no IBICT 1.2 Planejar uma revisão sistemática da literatura (RSL) sobre soluções de extração de textos utilizando técnicas de Machine Learning 1.3 Estudos para criação de um protocolo de revisão 1.4 Executar a Revisão de Literatura através do protocolo da revisão gerado 1.5 Selecionar e avaliar, principais estudos primários, que serão usados como fonte da pesquisa	Relatório Técnico	Até 8 meses
2. Organizar e desenvolver modelo baseado na IA para o processo de transcrição de imagens em texto	2.1 Estudos para verificação e Avaliação dos manuscritos que serão usados na pesquisa, quanto a época e tipo 2.2 Estudos para definição de um pequeno conjunto de manuscritos de diferentes épocas e tipos, que serão utilizados na pesquisa 2.3 Estudos para definição e desenvolvimento de uma API WEB para extração de palavras dos documentos escolhidos 2.4 Estudos para definição e desenvolvimento de API WEB de criação de rotinas para guarda em Banco de dados palavras exportadas 2.5 Validação do processo de extração de textos pelo Arquivo Nacional	Relatório Técnico	Até 10 meses
3. Organizar e aplicar modelo de	3.1 Estudos para definição e aplicação do modelo computacional	Relatório	Até

reconhecimento de texto em documentos manuscritos	processos de seleção dos manuscritos que serão transcritos 3.2 Estudos para definição e aplicação do modelo computacional baseada em técnicas de Inteligência Artificial definidas na revisão de literatura sistemática para transcrição dos manuscritos 3.3 Estudos para definição e aplicação do modelo computacional para visualização do texto transcrito	Técnico	13 meses
4. Validação do modelo computacional de correção e atualização por parte dos especialistas	4.1 Provar modelo computacional de processos para seleção de textos transcritos que serão avaliados pelos especialistas 4.2 Teste de correção amostral dos especialistas 4.3 Estudos para desenvolvimento e criação de modelo computacional de atualização e correções dos especialistas, possibilitando o aprendizado do modelo de Inteligência Artificial	Protótipo	Até 5 meses
5. Promover repasse de transferência das tecnologias para equipe do Arquivo Nacional	5.1 Levantamento das necessidades de treinamento da equipe do Arquivo Nacional 5.2 Workshops de compartilhamento de competências 5.3 Repasse de modelo computacional e tecnologia à equipe do Arquivo Nacional	Workshop	Até 7 meses
6. Disseminar e documentar todo o conhecimento da pesquisa	6.1 Elaboração de comunicação científica (artigos, pôsteres ou resumos para apresentação em congressos) 6.2 Oferta de treinamento à equipe de informação e técnica do Arquivo Nacional 6.3 Apresentação dos resultados da pesquisa	Artigos científicos / Treinamentos / Relatório Final	Até 12 meses

5. JUSTIFICATIVA E MOTIVAÇÃO PARA CELEBRAÇÃO DO TED

Nas últimas décadas, os investimentos para a transformação digital do material ligado ao patrimônio cultural vêm crescendo significativamente. Diversas Bibliotecas, Arquivos e outras instituições memoriais armazenam e disponibilizam grande número de livros e manuscritos históricos, que foram digitalizados para garantir sua preservação ao longo dos séculos.

O Arquivo Nacional é uma casa arquivística centenária. Nela encontram-se custodiados milhares de documentos físicos e digitais históricos e/ou probatórios para a sociedade. Nos últimos anos, um dos maiores desafios da instituição tem sido o de desenvolver projetos de digitalização do seu acervo físico.

Neste sentido, torna-se imprescindível o desenvolvimento de técnicas de preservação digital dos documentos, visando ampliar sua disseminação, facilitar seu acesso, promover sua conservação, gerar novos meios de utilização, dentre outras, favorecendo a gestão das coleções físicas no decorrer do tempo.

Cada vez mais, documentos físicos e manuscritos estão sendo digitalizados e disponibilizados para fins de estudos, análises e recuperação de textos. Muitos são registros históricos de uma nação, ou documentos probatórios que dão suporte às questões de direitos, por exemplo: tratados históricos, registros de entradas de imigrantes, registros de concessão de terras, registros de nascimento/óbito, etc. Muitos desses documentos, já digitalizados, encontram-se em sua forma bruta, escritos manualmente e, portanto, apresentam conteúdo não pesquisável.

Este fato prejudica as análises e leituras por parte do público interessado em seus conteúdos. Nos últimos anos, centenas de especialistas em paleografia, que trabalham na transcrição de documentos para o meio digital, vêm trabalhando em editores de textos convencionais, formato extremamente complexo e demorado. Alguns dos profissionais adotam softwares para suas transcrições automáticas que, embora agilizem o processo, acarretam muitos erros de grafia.

O presente projeto de pesquisa, visa preencher a lacuna existente entre a transcrição manual e a transcrição automática. A ideia é fomentar a pesquisa para o desenvolvimento de uma ferramenta que possibilite que imagens brutas sejam transcritas e analisadas pela inteligência artificial com capacidade de aprender e registrar todo o processo. Nessa conjuntura, espera-se que os primeiros resultados de transcrição automática obtidos sejam validados por especialistas em paleografia.

Este projeto é de grande importância para o Arquivo Nacional, pois auxiliará na otimização de estudo, desenvolvimento e implantação de diversas tecnologias que levariam muito tempo para serem pesquisadas e testadas sem o apoio de equipe habilitada para tal processo. As tecnologias aqui citadas vão ao encontro das necessidades de várias áreas da instituição, muitas dessas ainda com processos manuais ou softwares obsoleto.

A expertise desenvolvida pelo IBICT na área de Ciência da Informação está alinhada à demanda de pesquisa e de tecnologia levantada pelo Arquivo Nacional demonstra alinhamento com as competências da Coordenação-Geral de Tecnologias de Informação e Informática (CGTI) elencadas pelo Regimento Interno do Instituto (Portaria n. 3.443, de 10 de setembro de 2020), o qual declara, no art. 24, que a referida coordenação é responsável por

[...]

III - coordenar pesquisas orientadas à gestão da informação científica e tecnológica por meio de ativos computacionais e informacionais;

IV - coordenar pesquisas, no âmbito de competência do Instituto, tais como ciência de dados, tecnologias disruptivas, interoperabilidade de redes, apropriação de tecnologias, integração de sistemas, mecanismos de preservação da memória, dentre outras;

V - coordenar a implementação de projetos e construção de sistemas de informação, banco de dados e outros recursos computacionais produzidos, no âmbito do Instituto;

A partir dos aspectos apresentados, justifica-se a criação de projeto de pesquisa entre IBICT e Arquivo Nacional com o intuito de promover o estudo e o desenvolvimento de soluções a partir da pesquisa no âmbito conceitual, metodológico e tecnológico.

6. SUBDESCENTRALIZAÇÃO

A Unidade Descentralizadora autoriza a subdescentralização para outro órgão ou entidade da administração pública federal?

- ☐ Sim
- ☒ Não

7. FORMAS POSSÍVEIS DE EXECUÇÃO DOS CRÉDITOS ORÇAMENTÁRIOS

A forma de execução dos créditos orçamentários descentralizados poderá ser:

- ☐ Direta, por meio da utilização capacidade organizacional da Unidade Descentralizada.
- ☐ Contratação de particulares, observadas as normas para contratos da administração pública.
- ☒ Descentralizada, por meio da celebração de convênios, acordos, ajustes ou outros instrumentos congêneres, com entes federativos, entidades privadas sem fins lucrativos, organismos internacionais ou fundações de apoio regidas pela Lei nº 8.958, de 20 de dezembro de 1994.

8. CUSTOS INDIRETOS (ART. 8, §2º)

A Unidade Descentralizadora autoriza a realização de despesas com custos operacionais necessários à consecução do objeto do TED?

- ☒ Sim
- ☐ Não

O pagamento será destinado aos seguintes custos indiretos, até o limite de 20% do valor global pactuado:

1. Pagamento de Custos Operacionais e Administrativos de caráter indivisível, os quais serão utilizados para custear os gastos indivisíveis, usuais e necessários à consecução do objeto do contrato com a Fundação de Apoio autorizada do Ibict.

9. CRONOGRAMA FÍSICO-FINANCEIRO

Metas	Descrição	Unid. Medida	Qtde.	Valor Unit. (R\$)	Valor Total (R\$)	Início	Fim
1	Analisar e identificar bibliografia relacionada à Técnicas de Machine Learning	Un	1	R\$ 66.000,00	R\$ 66.000,00	Out. 2022	Mai. 2023
	- Habilitação da pesquisa no IBICT; - Planejar uma revisão sistemática da literatura (RSL) sobre soluções de extração de textos utilizando técnicas de Machine Learning; - Estudos para criação de um protocolo de revisão; - Executar a Revisão de Literatura através do protocolo da revisão gerado; - Selecionar e avaliar, principais estudos primários, que serão usados como fonte da pesquisa;	Relatório Técnico	-	-	-	-	-
2	Organizar e desenvolver modelo baseado na IA para o processo de transcrição de imagens em texto	Un	1	R\$ 99.600,00	R\$ 99.600,00	Out. 2022	Jul. 2023
	- Estudos para verificação e Avaliação dos manuscritos que serão usados na pesquisa, quanto a época e tipo; - Estudos para definição de um pequeno conjunto de manuscritos de diferentes épocas e tipos, que serão utilizados na pesquisa - Estudos para definição e desenvolvimento de uma API WEB para extração de palavras dos documentos escolhidos - Estudos para definição e desenvolvimento de API WEB de criação de rotinas para guarda em Banco de dados palavras exportadas - Validação do processo de extração de textos pelo Arquivo Nacional	Relatório Técnico	-	-	-	-	-
3	Organizar e aplicar modelo de reconhecimento de texto em documentos manuscritos	Un	1	R\$ 103.400,00	R\$ 103.400,00	Out. 2022	Nov. 2023
	- Estudos para definição e aplicação do modelo computacional processos de seleção dos manuscritos que serão transcritos - Estudos para definição e aplicação do modelo computacional baseada em técnicas de Inteligência Artificial definidas na revisão de literatura Relatório Técnico - Dez./ 2022 Abr./ 2023 sistemática para transcrição dos manuscritos - Estudos para definição e aplicação do modelo computacional para visualização do texto transcrito	Relatório Técnico	-	-	-	-	-
4	Validação do modelo computacional de correção e atualização por parte dos especialistas	Un	1	R\$115.500,00	R\$115.500,00	Out. 2023	Fev. 2024
	- Provar modelo computacional de processos para seleção de textos transcritos que serão avaliados pelos especialistas	Protótipo	-	-	-	-	-

	- Teste de correção amostral dos especialistas - Estudos para desenvolvimento e criação de modelo computacional de atualização e correções dos especialistas, possibilitando o aprendizado do modelo de Inteligência Artificial						
5	Promover repasse de transferência das tecnologias para equipe do Arquivo Nacional	Un	1	R\$31.900,00	R\$31.900,00	Mar. 2024	Set. 2024
	- Levantamento das necessidades de treinamento da equipe do Arquivo Nacional - Workshops de compartilhamento de competências - Repasse de modelo computacional e tecnologia à equipe do Arquivo Nacional	Workshop	-	-	-		-
6	Disseminar e documentar todo o conhecimento da pesquisa	Un	1	R\$82.158,21	R\$82.158,21	Set. 2023	Set. 2024
	- Elaboração de comunicação científica (artigos, pôsteres ou resumos para apresentação em congressos) - Oferta de treinamento à equipe de informação e técnica do Arquivo Nacional - Apresentação dos resultados da pesquisa	Artigos científicos Treinamentos Relatório Final	-	-	-	-	-

(*) Tanto o plano de trabalho, quanto o detalhamento orçamentário (plano de aplicação), para fins de transparência no processo, serão sempre revisados e atualizados, se necessário, durante a realização das diversas metas do projeto, a fim de incorporar informações adicionais coletadas no decorrer dos trabalhos, priorizar ações em decorrência de outras e para representar eventuais mudanças que possam surgir. Os custos indiretos referem-se à contratação da Fundação de Apoio para gestão financeira, conforme previsto no Decreto nº 10.426, de 16 de julho de 2020, e na Lei nº 8.958, de 20 de dezembro de 1994 e suas atualizações. Os valores de bolsa baseiam-se nas portarias Ibict nº 061/2014 e nº 021/2017, que dispõem sobre a contratação de bolsistas. Diárias e passagens são previsões para a participação em eventos técnicos-científicos da área, com a finalidade de apresentação dos resultados do projeto, primordiais na fase de disseminação dos modelos desenvolvidos. O documento apresenta a memória para cálculo do quantitativo de pesquisadores necessários à execução da pesquisa em relação ao tempo. A metodologia utilizada para definição do quantitativo e do perfil dos pesquisadores contemplou a definição de um modelo baseado em pesquisador padrão capaz de acompanhar toda a meta e atribuição de pesquisadores com diferentes perfis, conforme as atividades simultâneas previstas para cada meta. O quantitativo de pesquisadores e perfis também foi avaliado a partir da necessidade de se atuar de forma paralela para cumprimento dos objetivos do projeto. Por fim, o prazo de entrega das metas ainda foi utilizado como parâmetro para alocação de pesquisadores por períodos específicos. A definição de STPI ou STPF foi estabelecida para atividades previstas que não condizem com a execução de pesquisa, a exemplo da prestação de serviços de TI, diagramação de guias, cartilhas e livros, gráfica e outros.

10. CRONOGRAMA DE DESEMBOLSO

MÊS/ANO	VALOR*
OUT/2022	R\$498.558,21

(*) O valor contempla o recurso para execução da pesquisa e o custo indireto da fundação de apoio conforme detalhado no item 8 e explicitado no item 11.

11. PLANO DE APLICAÇÃO CONSOLIDADO - PAD

CÓDIGO DA NATUREZA DA DESPESA	CUSTO INDIRETO	VALOR PREVISTO
33.90.39	Não	R\$450.889,28
33.90.39	Sim	R\$47.668,93

12. PROPOSIÇÃO

Pela Unidade Descentralizada:

CECÍLIA LEITE OLIVEIRA
Diretora
Instituto Brasileiro de Informação em Ciência e Tecnologia

13. APROVAÇÃO

Pela Unidade Descentralizadora:

MAXIMILIANO MARTINS DE FARIA
Coordenador-Geral de Tecnologia da Informação
Superintendência de Planejamento e Gestão
Arquivo Nacional

JOELSON DA SILVA
Coordenador-Geral de Administração
Superintendência de Planejamento e Gestão
Arquivo Nacional



Documento assinado eletronicamente por **Cecilia Leite Oliveira, Usuário Externo**, em 20/10/2022, às 15:30, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Joelson da Silva, Coordenador-Geral de Administração**, em 20/10/2022, às 16:33, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Maximiliano Martins de Faria, Coordenador-Geral de Tecnologia da Informação**, em 20/10/2022, às 17:11, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site <https://sei.arquivonacional.gov.br/autentica>, informando o código verificador **0226756** e o código CRC **A2D07ADC**.