Glossário da Ciência Aberta

Milena Ambrosio Telles

Brasília, 27 de novembro de 2019.

Embrapa

Roteiro

- Contexto.
- Metodologia.
- Etapas já executadas.
- Fase atual.
- Próximas etapas.

O projeto

Atividade do Marco 4 - "Promoção de ações de sensibilização, participação e capacitação em Ciência Aberta" do **Compromisso 3** - “Estabelecer mecanismos de governança de dados científicos para o avanço da Ciência Aberta no Brasil” do **4º Plano de Ação Nacional para Governo Aberto (OGP).**

Equipe: Milena Telles (coordenadora do Glossário sobre Ciência Aberta) – Embrapa - milena.telles@embrapa.br, Maria Carolina Coradini, Eder Coimbra, Pedro Turci (alunos do 8º semestre do curso de Linguística da UFSCar). Vanessa Jorge (Fiocruz), Luana Sales (Ibict), Tatiane Pacanaro (Capes) + colaboradores.

Embrapa

Metodologia em Processamento de Línguas Naturais (PLN)

(Di Felippo, Almeida, 2010)

DOMÍNIO LINGUÍSTICO

- Seleção e delimitação do domínio especializado.
- Delimitação do conhecimento.
- Seleção das fontes e estratégia de aquisição do conhecimento.
- Extração do conhecimento léxico-conceitual.

DOMÍNIO REPRESENTACIONAL

- Representação do conhecimento em formato de ontologia, ou hierarquia conceitual.

DOMÍNIO IMPLEMENTACIONAL

- Codificação do conhecimento formalizado em uma base de dados.

Domínio linguístico

1. Seleção e delimitação do domínio especializado: *Ciência aberta.*
2. Delimitação do conhecimento:
    a. termos das categorias: substantivos (ou sintagmas nominais);
    b. lexias simples e compostas (1-grama); lexias complexas (2, 3, 4, 5, 6, 7, 8-gramas);
    c. *synsets* e relações conceituais;
3. Seleção das fontes e estratégia de aquisição do conhecimento:
    a. compilação do corpus utilizando expressão de busca para extrair artigos de periódicos especializados da área, teses, dissertações e textos do *google* geral;
4. Extração do conhecimento léxico-conceitual:
    a. listas de candidatos a termos a partir do corte de frequência delimitado pelo tamanho do *corpus* em número de palavras/100.000+1;
    b. validação e classificação dos termos por especialistas da área em subtemas para elaboração da ontologia.

1. Compilação, seleção, balanceamento e limpeza do *corpus*

Compilação de textos por expressão de busca

Obs.: resultado indireto - biblioteca no Zotero

Expressões de busca

Pilar	
Ciência aberta	Ciência Colaborativa, Pesquisa científica colaborativa, Conhecimento coletivo, Ciência compartilhada, Ciência colaborativa, Boas práticas de pesquisa, Ciência democrática, Acesso à ciência
Acesso aberto	-
Bloco de notas abertos	Cadernos de laboratórios abertos, cadernetas de campo abertas, cadernos de pesquisa abertos, dispositivos eletrônicos como bloco de notas, práticas em cadernos de pesquisa abertos, laboratórios compartilhados, laboratórios colaborativos
Dados abertos	Dados de pesquisa abertos, dados governamentais abertos, Acesso à ciência
Revisão por pares aberta	-
Ciência cidadã	Rede cidadã, Pesquisa e inovação social, Ciência democrática, Acesso à ciência

Pilar	
Código aberto	Projetos abertos de software livre, Licença pública, ferramentas, ferramentas livres
Fluxo de pesquisa aberto	Fluxo de pesquisa compartilhado
Recursos educacionais abertos	Educação Aberta, Rede de educação aberta, universidade aberta
Redes sociais científicas	
Repositórios científicos de acesso aberto	Repositórios institucionais, repositórios de dados de pesquisa, repositórios de pesquisa
Inovação aberta	-
Ética na ciência aberta	-
Propriedade intelectual e ciência Aberta	-

Expressões de busca

(ab("open access" OR "open data" OR "open science" OR "open research data" OR "open research" OR "research data" OR "open archives" OR "citizen science" OR "open hardware" OR "open software" OR "open educational resources") AND ab("citizen engagement" OR "public engagement" OR corruption OR discoverable OR "open development" OR "open government licence" OR "open movement" OR "transparency" OR "open advocacy" OR "open government data" OR "government data" OR "open government partnership" OR OGP OR data.gov OR "sharing data" OR "open access policies" OR collaboration OR reuse OR "use and reuse" OR "research data management" OR "data management planning" OR "open material" OR "reproducible research" OR "article processing charge" OR apc OR "open data" OR "share data" OR "data journal" OR "data journals" OR metadata OR "big data" OR "data mining" OR "text and data mining" OR "data analysis" OR "data standards" OR anonymization OR ontology OR taxonomy OR "hybrid journal" OR "hybrid model" OR "hybrid open" OR "gold open access" OR "green open access" OR "green route" OR "gold route" OR "diamond open access" OR "platinum open access" OR "open license" OR "creative commons" OR copyright OR embargo OR "intellectual property" OR license OR "immediate deposit" OR "optional access" OR "predatory publisher" OR reproducibility OR irreproducibility OR metric* OR "impact factor" OR "research impact" OR "research funder" OR "research funders" OR "funders policies" OR "governamental funders" OR "institucional funders" OR altmetrics OR bibliometrics OR semantometrics OR accessible OR discoverable OR fair OR foster OR openair OR opendoar OR "oai-protocol" OR pre-print OR preprint OR post-print OR "open peer review" OR "peer review" OR "open source" OR "linked data" OR "open lab notebooks" OR "open notebooks" OR "digital object identifier" OR DOI OR "persistent identifier" OR pid OR "long-term preservation" OR "self-archiving" OR depositing OR interoperability OR "institutional repository" OR "subject repository" OR "data repositories" OR "reporting bias" OR arXiv) AND PEER(yes) AND stype.exact("Scholarly Journals") AND at.exact("Book Chapter" OR "Reference Document" OR "Editorial" OR "Conference Paper" OR "Book" OR "Conference" OR "Article") AND la.exact("Portuguese" OR "Spanish" OR "English")) OR (ab("open access" OR "open data" OR "open science" OR "open research data" OR "open research" OR "research data" OR "open archives" OR "citizen science" OR "open hardware" OR "open software" OR "open educational resources") AND ab(glossary OR concept* OR iniciative* OR guideline* OR principle* OR standard* OR workflow* OR tool* OR policies OR "horizon 2020") AND PEER(yes) AND stype.exact("Scholarly Journals") AND at.exact("Book Chapter" OR "Reference Document" OR "Editorial" OR "Conference Paper" OR "Book" OR "Conference" OR "Article") AND la.exact("Portuguese" OR "Spanish" OR "English"))

Pastas organizadas no Zotero			Nº de arquivos	Nº de palavras
BRAPCI e literatura em CI			361	2.849.508
Teses e dissertações			35	1.720.749
Google acadêmico			49	645.795
Não científicos	Blog post		50	80.905
	Web page		50	306.545
	Journal articles	Apenas abstract	3	932
		Outros	7	27.474
		Artigos científicos	43	268.624
	Encyclopedia articles		8	14.366
	Conference papers		4	20.058
	Books		1	57.643
	Attachments	Artigos científicos	5	33.398
		Outros attachments	6	73.620
TOTAL			1.072	6.099.617

2. Seleção dos textos compilados e balanceamento do corpus

Corpus selecionado para extração de termos			
Pastas organizadas no Zotero		**N° de palavras (sujo)**	**Percentual de palavras**
Científicos (4.685.543; 85%)	**BRAPCI e literatura em CI**	**2.849.508**	**51,8%**
	Teses e dissertações + Books (1 Dissertação)	**1.778.392 + 57.643**	**32,4% + 1,0% = 33,4%**
Não científicos (807.633; 15%)	**Blog post**	**87.682**	**1,6%**
	Web page	**558.445**	**10,2%**
	Encyclopedia articles	**14.366**	**0,3%**
	Attachments: *Outros attachments*	**73.620**	**1,3%**
TOTAL		**5.493.276**	**100%**

3. Limpeza semiautomática do *corpus* com expressões regulares em Notepad++

Corpus selecionado para extração de termos			
Pastas organizadas no Zotero		**N° de palavras (limpo)**	**Percentual de palavras**
Científicos (3.705.115; 83,1%)	**BRAPCI e literatura em CI**	**2.238.415**	**50,2%**
	Teses e dissertações + Books (1 Dissertação)	**1.414.389 + 52.911**	**31,7% + 1,2% = 32,9%**
Não científicos (757.391; 15,7%)	**Blog post**	**76.486**	**1,7%**
	Web page	**549.000**	**12,3%**
	Encyclopedia articles	**11.427**	**0,3%**
	Attachments: *Outros attachments*	**60.239**	**1,3%**
TOTAL		**4.462.506**	**100%**

Extração do conhecimento léxico-conceitual:
listas de candidatos a termos, validação e classificação

Uso do AntConc, com corte de frequência 45:

N-grama	Nº de candidatos a termos extraídos	Pós limpeza	Pós validação
1-grama	6.445	88	
2-grama	11.634	148	
3-grama	4.010	135	
4-grama	804	56	
5-grama	188	14	
6-grama	61	4	
7-grama	3	1	
8-grama	2	1	
TOTAL	**23.147**	**447**	**6 esp. – 36 termos** **5 esp. – 52 termos** **88 termos validados**

2-gramas

Frequência	n° textos	Lexias complexas	Validação		Subtema	Observações
			Sim	Não		
71	4	acervo digital				
5703	9	acesso aberto				
96	7	acesso gratuito				
1375	9	acesso livre				
227	8	acesso público				
165	8	acesso restrito				
84	6	acesso universal				
586	8	administração pública				
219	6	ambiente digital				

3-gramas

Frequência	n° textos	Lexias complexas	Validação		Subtema	Observações
			Sim	Não		
511	7	abertura de dados				Synset 1
1098	8	acesso à informação				Synset 2
149	8	acesso à internet				
59	5	acesso à produção				
45	6	acesso aberto dourado				
182	8	acesso ao conhecimento				
56	7	acesso ao conteúdo				
375	7	acesso aos dados				
136	5	análise de conteúdo				
207	7	análise de dados				Synset 3
65	6	armazenamento de dados				
74	6	avaliação por pares				
676	8	banco de dados				
54	5	base de conhecimento				
926	9	base de dados				
119	1	bens comuns intelectuais				

4-gramas

Frequência	n° textos	Lexias complexas	Validação		Subtema	Observações
			Sim	Não		
50	5	access to research data				
102	7	acesso à informação científica				
63	6	acesso à informação pública				
70	8	acesso aberto à literatura				
55	8	acesso aberto a publicações				
79	7	acesso aberto ao conhecimento				
73	6	acesso aberto aos dados				
139	4	acesso livre à informação				
68	1	atividade de inovação aberta				
70	3	biblioteca de ensino superior				
75	8	budapest open access initiative				
48	2	caderno aberto de laboratório				
53	2	cadernos eletrônicos de laboratório				singular?

5-gramas

Frequência	n° textos	Lexias complexas	Validação		Subtema	Observações
			Sim	Não		
153	6	acesso aberto à informação científica				
94	4	acesso livre à informação científica				
48	4	biblioteconomia e ciência da informação				
50	5	ciclo de vida da curadoria				
235	7	ciclo de vida dos dados				
75	5	coleção de dados de pesquisa				
54	4	curadoria de dados de pesquisa				
49	8	directory of open access journals				
64	6	direito de acesso à informação				
326	7	gestão de dados de pesquisa				

Planilha para validação e classificação dos candidatos a termos por especialistas

A	B	C	D	E
Lexias	**Validação**		**Subtema**	**Observações**
	Sim	**Não**		
"exemplo"		x	clique na seta à direita >>>	"Escreva neste campo sugestões de sinônimos, que podem ou não estar presentes nessa lista; sugestão(ões) de subtema, se houver, quando marcar o subtema 'outros'; e outras observações que considerar pertinentes"
abertura de dados				
access to research data				
acervo				
acervo digital				
acesso				
acesso à informação científica				
acesso à informação pública				
acesso aberto a dados de pesquisa				
acesso aberto à informação científica				
acesso aberto à literatura				
acesso aberto a publicações				
acesso aberto ao conhecimento				
acesso aberto aos dados				
acesso aberto dourado				
acesso ao conhecimento				

- acesso aberto
- cadernos abertos de laboratório
- ciência cidadã
- código aberto
- dados abertos
- recursos educacionais abertos
- redes sociais científicas
- revisão por pares aberta
- outros

Exemplo de termos validados e classificados por especialistas

abertura de dados	x	x	x	x	x	x	6	dados	dados	acesso aberto	dados	dados
aprendizagem colaborativa	x	x	x	x	x	x	6	recursos	recursos	recursos	recursos	recursos
budapest open access initiative	x	x	x	x	x	x	6	acesso aberto	outros		acesso aberto	acesso aberto
caderno aberto de laboratório	x	x	x	x	x	x	6	cadernos	cadernos	cadernos	cadernos	cadernos
ciclo de vida dos dados	x	x	x	x	x	x	6	dados	dados	dados	dados	
ciência aberta	x	x	x	x	x	x	6	outros	outros	outros	outros	acesso aberto
ciência cidadã	x	x	x	x	x	x	6	ciência	ciência	ciência	ciência	ciência
código aberto	x	x	x	x	x	x	6	código aberto	código aberto	código aberto	código aberto	código aberto
colaboração científica	x	x	x	x	x	x	6	redes sociais	outros	redes sociais	redes sociais	ciência
compartilhamento de dados	x	x	x	x	x	x	6	dados	dados	dados	dados	cadernos
creative commons	x	x	x	x	x	x	6		acesso aberto	Tenho dúvida	dados	recursos
curadoria de dados de pesquisa	x	x	x	x	x	x	6	dados	dados	dados	dados	dados
dados abertos conectados	x	x	x	x	x	x	6	dados	dados	dados	dados	
declaração de berlim	x	x	x	x	x	x	6	acesso aberto	acesso aberto	acesso aberto	acesso aberto	acesso aberto
declaração de budapeste	x	x	x	x	x	x	6	acesso aberto	acesso aberto	acesso aberto	acesso aberto	acesso aberto
directory of open access journals	x	x	x	x	x	x	6	acesso aberto	acesso aberto	acesso aberto	acesso aberto	
educação aberta	x	x	x	x	x	x	6	recursos	recursos	recursos	recursos	recursos
gestão de dados	x	x	x	x	x	x	6	dados	dados	dados	dados	dados
hardware aberto	x	x	x	x	x	x	6	código aberto	código aberto	acesso aberto	código aberto	ciência
inovação aberta	x	x	x	x	x	x	6	código aberto	outros	ciência	redes sociais	ciência
laboratório cidadão	x	x	x	x	x	x	6	ciência	cadernos	ciência	ciência	ciência

Representação do domínio em ontologia – fase atual

Pontos de partida

- Classificação feita pelos especialistas a partir dos subtemas do “guarda-chuva” da Ciência Aberta.
- Estrutura conceitual do domínio na literatura mundial – VOSViewer.
- Categorias do projeto GovIE – Embrapa, que propôs modelo de Governança de dados e informações para a instituição.

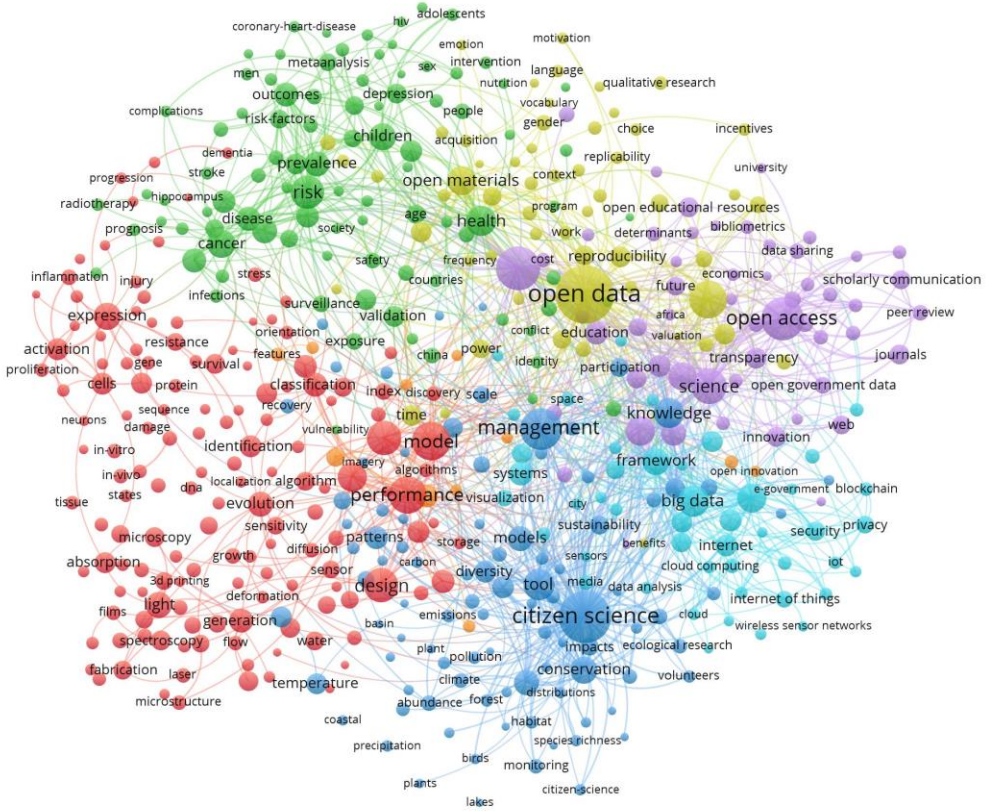
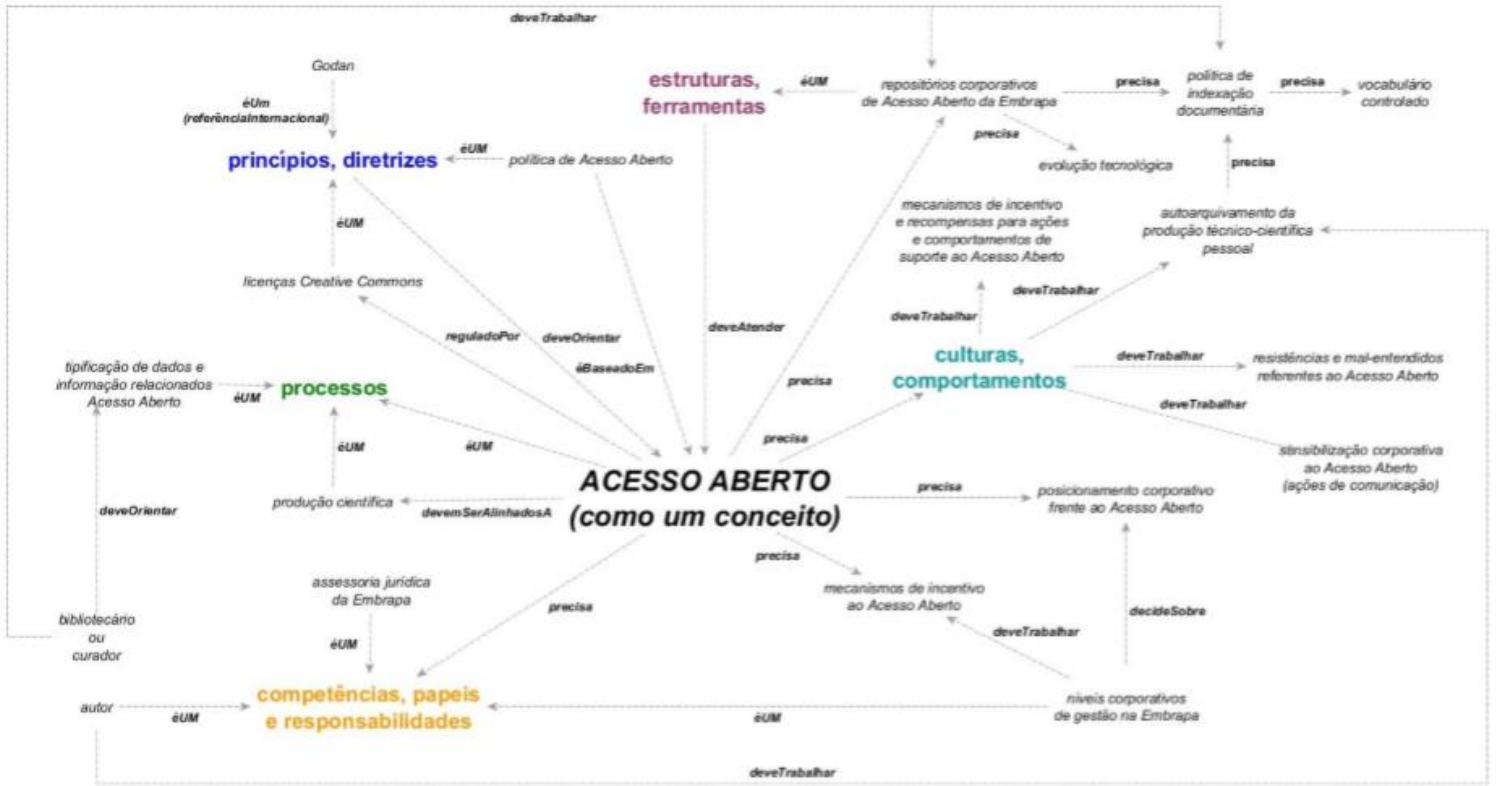

adolescents
hiv
coronary-heart-disease
emotion
motivation
metaanalysis
intervention
men
sex
language
nutrition
qualitative research
outcomes
depression
vocabulary
complications
people
risk-factors
gender
choice
incentives
children
acquisition
dementia
replicability
prevalence
university
progression
stroke
open materials
context
risk
radiotherapy
hippocampus
program
open educational resources
age
health
disease
work
bibliometrics
prognosis
society
determinants
cancer
data sharing
safety
frequency
cost
reproducibility
inflammation
injury
stress
countries
economics
scholarly communication
open data
future
infections
surveillance
validation
africa
peer review
expression
open access
orientation
exposure
conflict
education
valuation
activation
resistance
china
power
transparency
journals
proliferation
gene
survival
features
identity
participation
cells
protein
classification
index
discovery
scale
science
open government data
space
sequence
recovery
knowledge
neurons
damage
time
vulnerability
management
web
identification
model
innovation
in-vitro
imagery
algorithms
framework
in-vivo
localization
algorithm
systems
open innovation
dna
e-government
blockchain
states
tissue
evolution
performance
visualization
big data
city
microscopy
sensitivity
patterns
storage
models
sustainability
security
privacy
absorption
growth
diffusion
benefits
internet
carbon
sensors
cloud computing
iot
sensor
3d printing
design
diversity
tool
media
data analysis
deformation
internet of things
light
films
generation
emissions
citizen science
cloud
basin
wireless sensor networks
spectroscopy
flow
water
plant
impacts
ecological research
pollution
fabrication
laser
conservation
volunteers
temperature
climate
distributions
microstructure
abundance
forest
coastal
habitat
precipitation
species richness
birds
monitoring
plants
citizenscience
lakes
VOSviewer

Figura 9. Proposta de governança para aprimoramento da gestão da informação em suporte ao 'Acesso Aberto ao conhecimento científico na Embrapa' (Questão Prioritária no 4, Documento GovIE no. 1).



Fonte: Relatório do projeto especial GovIE, 2017.

Próximos passos

- Finalização da ontologia.
- Seleção dos termos a serem definidos.
- Busca por contextos definitórios (já inciada).
- Redação das definições.
- Validação das definições pelo grupo de especialistas.
- Publicação.

Obrigada!

milena.telles@embrapa.br

Embrapa