Brasília, 14 de fevereiro de 2022

## Campos de extração e padronização utilizada (scraping)

O conjunto de sites contém informações acerca de oportunidades de captação de recursos não-orçamentários provenientes de fundos de investimentos. No processo atual, a responsável tem que entrar periodicamente em mais de 100 sites e realizar buscas manuais de oportunidades de investimentos. O objetivo é substituir o trabalho manual por um automatizado, que colete periodicamente informações necessárias dos campos abaixo descritos, diretamente nos sites selecionados, e apresente as informações com alto nível de integridade.

As informações a seguir definem os campos que deverão ser raspados automaticamente dos sites para alimentar o arquivo CSV. O código em Python deverá buscar essas informações para alimentar o arquivo uma vez por dia e comparar se houve alteração com o dia anterior. É interessante gerar uma espécie de log para o caso de mudança na estrutura do site ou de algum tipo de erro de captura para possibilitar posterior análise e manutenção quando em produção.

São 4 campos de extração: oportunidades, notícias, política de financiamento e lista de projetos apoiados. A maior prioridade à época foi o campo oportunidades de financiamento.

Em cada um desses tópicos, é necessária a coleta de informações específicas (por meio de palavras-chave) para alimentar o CSV, informando se houve alteração nos dados coletados nos sites em relação às informações do dia anterior.

**Tabela 1. Oportunidades abertas (Grants, Fellowships, Scholarships)**

| **Elemento** | **Padronização de saída** |
|---|---|
| Nome da oportunidade | opo_titulo |
| Link da oportunidade | link |
| Deadline (se houver) | opo_deadline |
| Informação/texto da oportunidade (se houver) | opo_texto |
| Tipo de oportunidade: Classificado em (grant, scholarship, fellowship, other) em minúscula (caso não encontre os termos, classificar como 'other') | opo_tipo |
| Elegibilidade (busca de palavras-chave como brazil, latin américa etc) | opo_brazil |
| Código indexador ou ID pra cada item de toda a extração (comando padrão do Python definido no código exemplo) | codigo |
| Data da atualização daquela oportunidade (formato yymmdd) | atualizacao |

**Tabela 2. Notícias**

| **Elemento** | **Padronização de saída** |
|---|---|
| Título da notícia | not_titulo |
| Link da notícia | link |
| Descrição da notícia | not_texto |
| Código indexador ou ID pra cada item de toda a extração | codigo |
| Data da atualização daquela oportunidade (formato yymmdd) | atualizacao |

**Tabela 3. Políticas de financiamento**

| Elemento | Padronização de saída |
|---|---|
| Nome do site proveniente da política | pol_instituicao |
| Nome do campo de extração (about us, how we work...) | pol_titulo |
| Designação da política | pol_texto |
| Link da página | link |
| Código indexador ou ID pra cada item de toda a extração | codigo |
| Data de atualização daquela oportunidade (formato yymmdd) | atualizacao |

**Tabela 4. Lista de projetos apoiados pelo fundo**

| Elemento | Padronização de saída |
|---|---|
| Título do projeto | prj_titulo |
| Link do projeto | link |
| Instituição (se houver) | prj_instituicao |
| Valor (se houver) | prj_valor |
| Se contém Brasil, América Latina, América do Sul etc | prj_brazil |
| Código indexador ou ID pra cada item de toda a extração | codigo |
| Data da atualização daquela oportunidade (formato yymmdd) | atualizacao |