# Proposta de um *Dataset* para a Agricultura Utilizando Dicionário de Termos *Agrovoc* como Fonte de Dados

**Clovis S. Junior**[1], **Carina F. Dorneles**[2]

[1]Instituto de Ciências Exatas e Naturais/UFR
Rondonópolis, MT/Brasil

[2]Departamento de Informática e Estatística - INE
Universidade Federal de Santa Catarina - UFSC/Florianópolis

clovis@ufr.edu.br, carina.dorneles@ufsc.br

***Abstract.*** *Scientific production largely depends on specific technical terms for each area of knowledge. These terms usually do not follow a conventional vocabulary or free translation, especially in sharing data and information between languages. In the agricultural domain, the Food and Agriculture Organization of the United Nations (FAO) contributes with an important tool to help the information exchange in different languages: the Adrovoc dictionary. Agrovoc contributes as a vocabulary of terms, also composed of a collaborative ontology with terms in up to 40 languages. The availability of this vocabulary together with an ontology can make it challenging to integrate with third-party systems such as Enterprise Resource Planning or ERP. The contribution of this work is to present an alternative for creating a tabular dataset from the Agrovoc so that the terms could be used in relational databases, allowing easy integration with applications for agriculture.*

***Resumo.*** *A produção científica depende em grande parte de termos técnicos específicos para cada área do conhecimento. Esses termos usualmente não seguem um vocabulário convencional ou tradução livre principalmente no compartilhamento de dados e informações entre idiomas. No domínio agrícola a Organização das Nações Unidas para Alimentação e Agricultura (FAO) contribui com uma importante ferramenta para auxiliar o intercâmbio de informações em diferentes idiomas: o dicionário Agrovoc. O Agrovoc funciona como um vocabulário de termos composto também por uma ontologia colaborativa contando com termos em até 40 idiomas. A disponibilização desse vocabulário juntamente com a estrutura de uma ontologia pode dificultar a integração com sistemas de terceiros, como por exemplo Enterprise Resource Planning ou ERP. A contribuição desse trabalho está na apresentação de uma alternativa para criação de um dataset tabular a partir do dicionário Agrovoc para uso dos termos em bases de dados relacionais permitindo facilmente a integração com aplicações destinadas à agricultura.*

## 1. Introdução

A Organização das Nações Unidas para Alimentação e Agricultura (FAO) desenvolveu um dicionário de termos chamado *Agrovoc*[1] como referência para trabalhos técnicos,

[1]https://agrovoc.fao.org/

acadêmicos e demandas relacionadas à tradução de termos equivalentes entre idiomas. A estrutura do *Agrovoc* utiliza *Resource Description Framework* (*RDF*).

Alguns exemplos para uso do Agrovoc podem ser observados na pesquisa desenvolvida por [Aryal et al. 2014] com o framework GEOBIA, nessa proposta é explorada a criação de objetos de imagem a partir de capturas de imagens espacial com alta resolução. Em geral, os objetos de imagem extraídos são prontamente utilizados no formato vetorial pronto para sistemas de GIS. Nesta pesquisa, os autores desenvolveram uma semântica temática e espacial usando o AGROVOC para extrair vocabulário para o domínio agrícola. Os resultados mostraram que no framework GEOBIA, os objetos podem ser caracterizados com significado semântico e sua relação com o mundo real.

Outro exemplo para uso do Agrovoc é apresentado por [Beneventano et al. 2013] com o banco de dados CEREALAB para auxiliar criadores na escolha de marcadores moleculares associados às características mais importantes. Os dados fenotípicos e genotípicos são obtidos a partir da integração de bancos de dados de código aberto.

A contribuição apresentada neste artigo cria um *dataset* relacional, com os termos agrícolas aceitos pela comunidade científica internacional. O propósito de ser utilizado como referência na tradução e consulta de termos equivalentes em idiomas diferentes. Nesse sentido, a extração e importação de dados para bases de dados relacional armazenado localmente se justifica por 2 razões: 1) Disponibilidade: dados externos em *datasets* públicos eventualmente sofrem evoluções ou descontinuidades. Dessa forma, *datasets* tem disponibilidades cronologicamente incertas sem o controle de pesquisadores, estudantes e outros interessados. 2) Organização: tanto o formato original dos *datasets* quanto a organização interna dos dados são importantes para mais viabilidade em utilizá-los. Formatos proprietários de difícil compreensão ou estruturas com excesso de dados dificultam o uso de *datasets* relevantes para usuárioe finais com puca familiaridade no uso de *web* semântica. A importação de dados também proporciona a limpeza de dados desnecessários, garantindo a viabilidade do uso dos dados.

A contribuição deste artigo é apresentar um facilitador por meio de um *dataset* simples com acesso rápido a termos específicos da área agrícola para auxiliar tanto pesquisadores quanto pessoas interessadas em desenvolver trabalhos com demandas para tradução.

Este artigo está organizado como segue. Na Seção 2 são descritos os trabalhos relacionados, usados para contextualizar os conceitos de metadados e ontologias. A Seção 3 presenta uma visão geral a respeito do dicionário de termos utilizado como fonte de dados para a criação do *dataset* proposto. Na Seção 4, são apresentadas: parser e estrutura de banco de dados proposta o desenvolvimento do *dataset* para o domínio do agronegócio. O *dataset* resultante e os passos utilizados para criá-lo são apresentados na Seção 5. Na Seção 6, são apresentadas as conclusões e discutidos alguns trabalhos futuros.

## 2. Fundamentação Teórica

Nesta seção serão abordados alguns conceitos citados e pertinentes à proposta apresentada no artigo. A versão atual do dataset proposto contém somente os termos principais extraídos da ontologia agrovoc, os demais dados como conceitos e classificações para os termos estão em fase de enriquecimento por meio de extração.

Ontologias fornecem uma forma de compartilhamento e reuso do conhecimento a respeito de um domínio específico, esse conceito tem sido aplicado em diversas áreas, tais como aplicações com Web Semântica, *e-commerce*, medicina, indústria automotiva, gestão financeira ou outra área relacionada com a gestão de conhecimento [Arp et al. 2015]. Ontologias definem os termos utilizados para descrever e representar uma área de conhecimento e podem ser compartilhados por pessoas, banco de dados e aplicações específicas. Ontologias também têm sido utilizadas para descrever estruturas de dados de forma hierárquica [Dermeval et al. 2015]. A partir de uma ontologia, é possível criar um padrão para fornecer metadados para a descrição e intercâmbio de dados entre os elementos envolvidos nos processos de um domínio específico.

Segundo [NISO 2022], metadados são informações estruturadas que descrevem, explicam, localizam ou tornam mais fácil a recuperação de dados, melhorando o gerenciamento de recursos e informações. Metadados são muitas vezes chamados de dados a respeito de dados. O termo metadado é usado de forma diferente dependendo da comunidade na qual se aplica. Alguns usam para se referir à informação compreensível por máquina, enquanto outros usam apenas para representar registros que descrevem recursos eletrônicos. A abordagem referente a metadados no contexto do artigo restringe-se a aplicações agrícolas, nesse sentido serão apresentados os padrões de metadados mais conhecidos para esse domínio:

- *Dublin Core*: Vocabulário composto por quinze propriedades usado para descrever recursos, é composto por quinze elementos que fazem parte de um conjunto maior de vocabulários de metadados e especificações técnicas mantidas pelo *Dublin Core Metadata Initiative - DCMI* [Maron and Feinberg 2018].
- *Darwin Core*: Padrão de metadados composto por um vocabulário de termos para facilitar a descoberta, recuperação e integração das informações sobre os organismos de qualquer natureza, verificando as sua ocorrência espaço-temporais e provas localizadas em coleções biológicas [Wieczorek et al. 2012].
- AgMES: A iniciativa da *Agricultural Metadata Element Set* ou simplesmente *AgMES* visa contemplar questões de semântica normas no domínio da agricultura com relação à descrição, a descoberta de recursos, a interoperabilidade e o intercâmbio de dados para os diferentes tipos de informação [eng Fao 2003].
- *AGRIS*: O *AGRIS Application Profile* é um padrão criado especificamente para melhorar a descrição, o intercâmbio e a posterior recuperação de dados referentes à produção agrícola com o *Documents-Like information Objects* (*DLIO*s) [Salokhe et al. 2005].
- *agXML*: Padrão desenvolvido para atender a formalização de dados no segmento agrícola de grãos favorecendo também o processamento de informações empresariais e entidades relacionadas [AgXML 2022].
- *agroXML*: O padrão fornece um método de armazenamento de dados estruturados por assuntos agrícolas. Os dados podem ser armazenados ou trocados entre os diferentes participantes e inclui elementos particulares da cultura com certa restrição [Martini et al. 2013].
- *Agrovoc*: Surgiu como um dicionário de sinônimos, mas atualmente está evoluindo para uma ontologia. Esse é um conceito novo que está surgindo em várias iniciativas na Web Semântica que pode ser definido como um sistema semântico que contém termos. Esse sistema semântico pode ser referido como

"Service Ontology". Criado e mantido pela *Food and Agriculture Organization* ou *FAO*, disponível desde o início dos anos 80, com atualizações constantes [Mietzsch et al. 2021, Simek et al. 2018].

## 3. *Agrovoc*

A cobertura do *Agrovoc* incluiu todas as áreas de interesse para a *Food and Agriculture Organization - FAO*, como agricultura, pesca, nutrição, silvicultura e meio ambiente, foi usado pela primeira vez na indexação *AGRIS* (*International Information System for the Agricultural Sciences and Technology*), como um banco de dados de domínio público global disponibilizado pela FAO, atualmente conta com aproximadamente três milhões de registros bibliográficos estruturados. Até o ano 2000, a FAO mantinha também uma versão impressa do Agrovoc, sendo esta migrada integralmente para mídia digital, com armazenamento feito por um banco de dados relacional disponibilizada no formato *Microsoft Access*, a versão relacional também foi descontinuada sendo esta migrada para uma ontologia[2]. Esta foi uma grande melhoria em termos de facilidade de manutenção. A Tabela 1 mostra um comparativo entre os termos de diversos dicionários e o *Agrovoc* [Caracciolo et al. 2013]. Apresentando de forma quantitativa as correspondências entre os mesmos.

| Vocabulário | Área | Idioma | Correspondências |
|---|---|---|---|
| EUROVOC | Geral | Inglês | 1,297 |
| DDC | Geral | Inglês | 409 |
| LCSH | Geral | Inglês | 1,093 |
| NALT | Agricultura | Inglês | 13,390 |
| RAMEAU | Geral (Agrícola) | Francês | 686 |
| DBpedia | Geral | Inglês | 1,099 |
| TheSoz | Ciências Sociais | Inglês | 846 |
| STW | Economia | Inglês | 1,136 |
| FAO Geopol. Ontology | Geopolítica | Inglês | 253 |
| GEMET | Meio Ambiente | Inglês | 1,191 |
| ASFA | Ciências Aquáticas | Inglês | 1,812 |
| Biotech | Biotecnologia | Inglês | 812 |
| GeoNames | Geografia | Inglês | 212 |

**Tabela 1. Recursos Associados ao *Agrovoc***

.

A Tabela 1 também mostra, para cada recurso vinculado ao *AGROVOC* (coluna1:Vocabulário), a área de abrangência (coluna2:Área), o idioma considerado para mapeamento com o *AGROVOC* (coluna3:Idioma) e o número de correspondências resultantes da avaliação (coluna4:Correspondências).

Segundo [Simek et al. 2018], o aumento constante do volume de dados, criando uma demanda para descrevê-los de forma eficiente, ou seja, fornecer principalmente uma descrição de conteúdo de qualidade e dispor de ferramentas para sua classificação, pesquisa, compartilhamento, administração, ou, conforme o caso, também para a sua

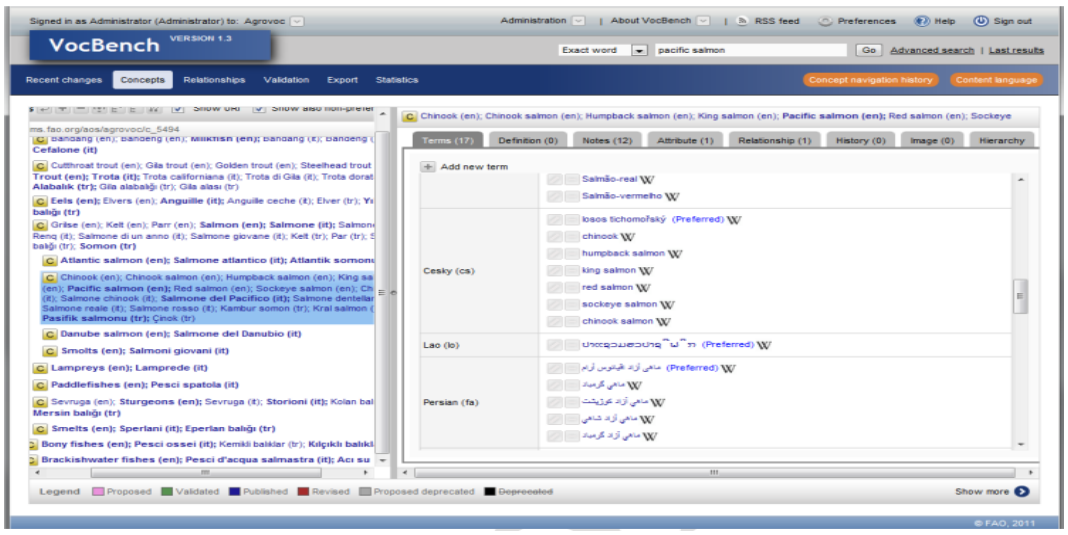
[2]https://agrovoc.fao.org/browse/agrovoc/en/

automatização de distribuição. Atualmente, essa tendência pode ser observada em todas as esferas da atividade humana, incluindo o setor agrário, no qual novos dados adquiridos não apenas por humanos, mas também de forma automatizada, resultando em grandes volumes de dados com necessidades de recursos para compartilhá-los.

Outro ponto importante refere-se à infraestrutura técnica do Agrovoc, baseada em um ecossistema abrangente de ferramentas para usuários e editores fornecendo acesso aos dados para humanos e máquinas. *Agrovoc* é um recurso estável e confiável, que é continuamente expandido pela atividade dos curadores e da comunidade editorial. Melhorias na tecnologia subjacente levaram a melhorias na representação de conteúdo. Os dados também são disponível em formatos legíveis por máquina como *RDF/XML*, *Turtle* e *JSON-LD* para consulta, seleção e extração de subconjuntos. Também está disponível a linguagem de consulta *SPARQL* para dados semânticos. Os resultados das consultas podem ser exibidos em tabelas e podem ser copiados como arquivos em vários formatos, incluindo valores separados por vírgula (*CSV*) [Mietzsch et al. 2021].

Alguns editores de ontologia, como *Protégé*[3] apesar de disponíveis desde o início do *Agrovoc*, não criaram extensões de requisitos do *Agrovoc*. Esse trabalho começou somente em 2004 com o conceito *Agrovoc Server Workbench*, também conhecido como "The WorkBench", um editor de vocabulário totalmente multilíngue baseado na web que suporta colaboração distribuída estruturada em um fluxo de trabalho. O sucessor dessa ferramenta é o *VocBench*, que possui melhorias em relação ao antecessor, pois suporta um fluxo de trabalho de edição formalizado por função de usuário e idioma, incluindo um mecanismo de rastreamento de alterações refinado que permite que indivíduos e organizações contribuam para o *Agrovoc* enquanto mantêm informações de proveniência relativas à autoria. A Figura 1 apresenta uma interface de usuário *VocBench* mostrando um fragmento do *Agrovoc*. Apesar da interface possuir recursos robustos para consulta aos termos, há a dependência de uma interface não integrada a outras aplicações, o presente artigo procurou atender essa demanda.

**Figura 1. VocBench v1.3 Fragmento de Interface para usuário com AGROVOC**




[3]https://protege.stanford.edu/

## 4. Proposta

A popularização e a facilidade no uso dos termos disponíveis no *Agrovoc* é um ponto central para esta proposta, visando proporcionar que usuários e sistemas não especializados tenham acesso ao conteúdo específico do dicionário de termos sem conhecimentos prévios em *Web* Semântica e também possam se beneficiar dos termos especificados, em diferentes idiomas.

A criação do *dataset* apresentado nesse artigo foi feita a partir da extração de um subconjunto de dados contidos no dicionário *Agrovoc* disponibilizado pela *FAO*. A metodologia consistiu essencialmente em verificar a estrutura do arquivo *RDF* disponível publicamente para identificar as *tags* correspondentes aos termos dos idiomas disponíveis. Definido o subconjunto de *tags* no conteúdo do arquivo *RDF* foi construída uma ferramenta para analisar o conteúdo identificando os termos associados ao respectivo idiomas. Posteriormente os termos foram utilizados para compor um *script* em *Structure Query Language - SQL* para criação da base de dados. Alguns termos não estão disponíveis em todos os idiomas compartilhados pelo *Agrovoc*, de qualquer forma todas os atributos com os dados disponíveis foram preenchidos totalizando mais de 40.000 linhas de dados no *dataset* final.

### 4.1. Construção do *Dataset*

Conforme apresentado no algoritmo da Figura 2 a construção do *dataset* foi feita através de um gerador de *script*, responsável pela extração dos dados formalizando-os em *SQL*. Para isso, foi desenvolvido um mini parser para identificar as *tags* que delimitam os blocos de dados referente aos termos e idiomas disponíveis, extraindo também os valores das *tags* para posteriormente serem utilizados no *script* resultante. A implementação do mini parser (*procedure*) foi desenvolvida para usar um arquivo contendo os somente os termos nos 40 idiomas disponíveis no *Agrovoc*, esse arquivo foi criado manualmente a partir de um editor de texto convencional mantendo somente as *tags* os termos disponíveis. O algoritmo utiliza esse arquivo como fonte para a pesquisa dos termos e limpeza de dados desnecessários para a criação do *script* em *SQL* utilizado posteriormente para gerar os dados em uma tabela relacional. A pesquisa verifica os termos delimitados em *tags RDF*, os termos seguem uma sequência em ordem alfabética com os idiomas disponíveis, o conteúdo de cada *tag* extraída é temporariamente armazenado em um vetor de textos, após o preenchimento de todos os elementos o algoritmo finaliza a construção sintática da sentença em *SQL* preenchendo os valores finais com os elementos contidos no vetor com cada um dos 40 idiomas, essa operação se repete até o final do arquivo contendo o subconjunto de termos utilizados para construção do *dataset*. É importante destacar que o arquivo de dados resultante após o processamento do algoritmo não é a base de dados final, o resultado é um *script SQL* que pode ser utilizado para o preenchimento da tabela com termos em diversos banco de dados, pois não há tipos de dados específicos, dessa forma o *script* torna-se genérico podendo ser utilizado em gerenciador de banco de dados diferentes. Termos sem correspondência em certos idiomas recebem um valor nulo (*null*) para o preenchimento possibilitado em consultas futuras a verificação dos idiomas disponíveis para o cada termo. A estratégia adotada para a criação do *dataset* utilizando um *script* intermediário justifica-se para proporcionar independência quanto a escolha do gerenciador de dados. Neste artigo, foi usado o banco de dados *SQLite*, entretanto, o *script* não é criado especificamente para esse banco de da-

dos, os tipos de dados utilizados são comuns a outros gerenciadores. O arquivo gerado com o mini parser contendo as sentenças *SQL* para a criação do *dataset* está disponível no repositório `https://github.com/clovissjunior/Agrovoc_Dataset/blob/main/Agrovoc_Terms.zip`, com tamanho físico: 38.014 MB

**Figura 2. Algoritmo Ilustrativo para Criação de Sentenças**

```
Procedimento ParserRdf_SQL
  Entrada, Saida : Arquivos
  Conteudo : Texto
  Início
    Entrada◄— Agrovoc.rdf
    Saida  ◄— Agrovoc.sql
    Enquanto !FinalArquivo(Entrada)
    Início
      Leia(Entrada,Conteudo)
      Se (Subtexto(Conteudo)='<RDF')
       Escreva(Saida, `INSERT INTO agrovoc_terms (
       arabic,           catalan, czech,      danish,   german,    english,    spanish,          estonian,
       persian,          finnish,  french,     hindi,    hungarian, indonesian, italian,          japanese,
       georgian,         korean,  latin,      laotian,  malay,     burmese,    norwegianbokmal, dutch,
       norwegiannynorsk, polish,  portuguese, romanian, russian,   slovak,     slovenian,        serbian,
       swedish,          swahili, telugu,     thai,     turkish,   ukrainian,  vietnamese,       chinese')
      Senão
       Se (Subtexto(Conteudo)='</RDF')
       Inicio
         Para Indice  1..40
           ValoresSentenca◄—ValorSentenca + VetorTextos[indice]
         Escreva(Saida,ValorSentenca)
       Fim
    Fim
 Fim
Senão
  Se (Conteudo) = ('ar','ca','cs','da','de','en','es','et','fa','fi',
                   'fr','hi','hu','id','it','ja','ka','ko','la','lo',
                   'ms','my','nb','nl','nn','pl','pt','ro','ru','sk',
                   'sl','sr','sv','sw','te','th','tr','uk','vi','zh') VetorTextos[]◄— Conteudo
Fim-Procedimento
```

## 4.2. Ferramenta utilizada

A criação do *dataset* foi realizada utilizando um parser escrito em *Object Pascal* com interface gráfica em *Lazarus*[4]. O algoritmo realizada pesquisas textuais em um subconjunto de dados da fonte original do dicionário (*RDF*), é uma implementação simples, o código completo referente a identificação das entidades e a extração dos valores das *tags* estão disponíveis no *github* no endereço `https://github.com/clovissjunior/Agrovoc_Dataset/blob/main/Agrovoc_Source_Code.pas` com acesso público. O algoritmo constrói o *script* (*Inserts-SQL*) para criação do *dataset*, optou-se por essa solução para proporcionar independência na escolha do banco de dados a ser utilizado, o *script* é genérico sendo compatível com os bancos de dados relacionais mais utilizados.

O *SQLite* foi o banco de dados utilizado para a criação da versão do *dataset* apresentado nesse artigo e estruturalmente consiste em uma única tabela. A proposta de criar uma estrutura única simplifica o uso por diferentes tipos de aplicações sem a necessidade de construir consultas mais complexas. A Figura 3 mostra a estrutura utilizada em *SQLite*.

A Figura 4 mostra um fragmento do subconjunto de dados do arquivo original utilizado para a representação do dicionário. O *dataset* resultante possui 16.877 MB de armazenamento físico com 40.257 linhas.

[4]https://www.lazarus-ide.org/

**Figura 3. DDL em SQLite para Criação do Dataset.**

```
[Agrovoc_Terms]
id         integer  PRIMARY key, norwegiannynorsk VARCHAR(70) NULL, norwegianbokmal  VARCHAR(70) NULL,
arabic     VARCHAR(70) NULL, catalan     VARCHAR(70) NULL, czech      VARCHAR(70) NULL, danish    VARCHAR(70) NULL,
german     VARCHAR(70) NULL, english     VARCHAR(70) NULL, spanish    VARCHAR(70) NULL, estonian  VARCHAR(70) NULL,
persian    VARCHAR(70) NULL, finnish     VARCHAR(70) NULL, french     VARCHAR(70) NULL, hindi     VARCHAR(70) NULL,
hungarian  VARCHAR(70) NULL, indonesian VARCHAR(70) NULL, italian    VARCHAR(70) NULL, japanese  VARCHAR(70) NULL,
georgian   VARCHAR(70) NULL, korean      VARCHAR(70) NULL, latin      VARCHAR(70) NULL, laotian   VARCHAR(70) NULL,
malay      VARCHAR(70) NULL, burmese     VARCHAR(70) NULL, dutch      VARCHAR(70) NULL, serbian   VARCHAR(70) NULL,
polish     VARCHAR(70) NULL, portuguese VARCHAR(70) NULL, romanian   VARCHAR(70) NULL, thai      VARCHAR(70) NULL,
russian    VARCHAR(70) NULL, slovak      VARCHAR(70) NULL, slovenian  VARCHAR(70) NULL, swedish   VARCHAR(70) NULL,
swahili    VARCHAR(70) NULL,telugu       VARCHAR(70) NULL,turkish     VARCHAR(70) NULL, ukrainian VARCHAR(70) NULL,
vietnamese VARCHAR(70) NULL,chinese      VARCHAR(70) NULL);
```

**Figura 4. Fragmento da Estrutura XML Derivada do Agrovoc**

```
<rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/c_4788">
  <skos:prefLabel xml:lang="ar">/>أساليبskos:prefLabel>
  <skos:prefLabel xml:lang="it">Metodi</skos:prefLabel>
  <skos:prefLabel xml:lang="pl">Metoda</skos:prefLabel>
  <skos:prefLabel xml:lang="ru">методы</skos:prefLabel>
  <skos:prefLabel xml:lang="sk">metódy</skos:prefLabel>
  <skos:prefLabel xml:lang="ms">Kaedah</skos:prefLabel>
  <skos:prefLabel xml:lang="tr">yöntem</skos:prefLabel>
  <skos:prefLabel xml:lang="uk">методи</skos:prefLabel>
  <skos:prefLabel xml:lang="ro">metode</skos:prefLabel>
  <skos:prefLabel xml:lang="pt">método</skos:prefLabel>
  <skos:prefLabel xml:lang="sr">методе</skos:prefLabel>
  <skos:prefLabel xml:lang="sw">mbinu</skos:prefLabel>
</rdf:Description>
```

## 5. Resultados

A Figura 5 mostra um fragmento do *dataset* final proposto no artigo, nesta ilustração é possível verificar a diversidade tanto de termos quanto de idiomas disponíveis no *Agrovoc*, é importante destacar que, dependendo do idioma pode haver a necessidade de ajustes na plataforma que se deseja utilizar, pois pode haver incompatibilidade nas tabelas de caracteres para visualização correta dos símbolos em alguns principalmente idiomas que utilizam alfabetos baseados em símbolos. A escolha do banco de dados *SQLite* levou em consideração esse requisito, pois nesse contexto, não há necessidade de configurações adicionais para a visualização de qualquer tipo de idioma. Os testes realizados com esse banco de dados mostraram-se viáveis tanto para a visualização com interfaces textuais quanto gráficas, mas eventuais ajustes podem ser necessários dependendo dos parâmetros ou configurações do sistema operacional ou a plataforma que se deseja utilizar.

**Figura 5. Fragmento do *Dataset* Final.**

| Arabic | German | Portuguese | English | Laotian |
|---|---|---|---|---|
| رعاية موسع | Extensivhaltung | criação extensiva | extensive husbandry | ການລ້ຽງສັດແບບທຳມະຊາດ |
| خلطات علفية | Feedlot | feedlot | feedlots | ການລ້ຽງສັດແບບຂັງ |
| جز | Schur | tosquia | shearing | ຕັດຂົນສັດ |
| صوف | Wolle | lã | wool | ຂົນແກະ |
| تشتية | Überwintern | invernação | wintering | ການລ້ຽງສັດໃນລະດູໜາວ |
| رش | Spritzen | pulverização | spraying | ການພົ່ນ |
| نثر (البذور) | Breitwurf | distribuição a lanço | broadcasting | ການຫວ່ານ |
| غمر | Tauchbad | imersão | dipping | ການຈຸ່ມ |
| نقع | Durchtränken | embebição | soaking | ການແຊ່ໃນທາດແຫຼວ |
| تعفير | Stäuben | polvilhação | dusting | ການພົ່ນແບບຜົງ |

### 5.1. Compartilhamento do *Dataset*

O propósito do *dataset* apresentado neste artigo é o mesmo do dicionário *Agrovoc*, compartilhamento de dados, entretanto procurou-se simplificar o acesso às aplicações por terceiros, dessa forma, a base de dados resumida pode facilmente ser integrada a qualquer sistema que necessite de consultas periódicas a termos agrícolas nos idiomas disponibilizados pelo *Agrovoc*.

A FAO disponibiliza o dicionário Agrovoc integralmente como ontologia, dessa forma, extrações de subconjuntos para demandas mais específicas tornam-se difíceis tanto do ponto de vista de implementação quanto processamento. Apesar do volume físico dos dados, cerca de 1.2 GB, não ser elevado, a estrutura da ontologia torna muito difícil a sua manipulação por aplicações relacionadas à extração e pesquisa por termos e conceitos.

Quanto ao público alvo para o *dataset* proposto, o cenário primário para utilização do *dataset* derivados do *Agrovoc* é o meio acadêmico. Neste contexto, constantemente faz-se uso de referências para consulta de termos técnicos principalmente na construção de pesquisas como artigos científicos ou mesmo na tradução de ferramentas de associadas à área de tecnologia da informação nos domínios relacionados com engenharia agrícola, engenharia ambiental, engenharia agronômica e áreas afins.

O conjunto de dados está disponível de forma pública no repositório github (`https://github.com/clovissjunior/Agrovoc_Dataset/blob/main/agrovoc.db`), sendo possível a realização de cópias sem restrições. Enfatizando que contribuições para a melhoria da estrutura do *dataset* são aceitas.

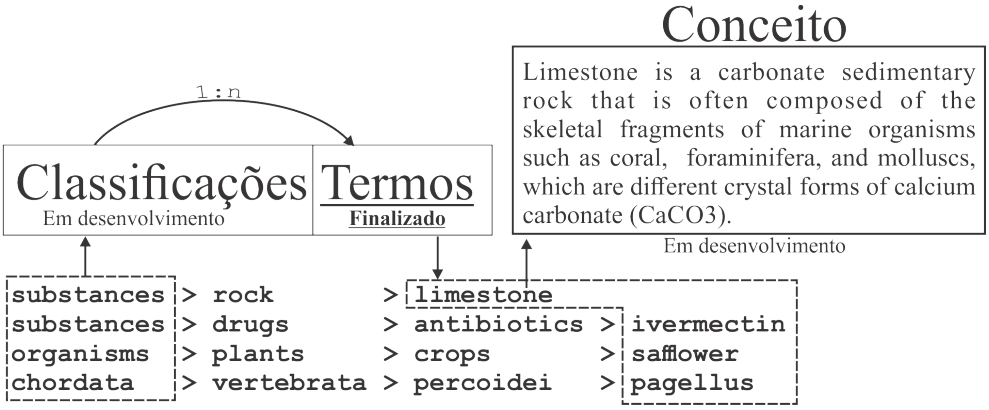
### 5.2. Base de dados Resultante

A base de dados resultantes no conjunto de dados refere-se à extração de um subconjunto do *Agrovoc*, com isso parte do arquivo original não foi utilizado, pois são dados referentes à hierarquia e definição de termos. O *script* para criação da estrutura do *dataset* está disponível publicamente no repositório `https://github.com/clovissjunior/Agrovoc_Dataset/blob/main/Script_Tabela_SQLite.sql`. A estrutura original do arquivo possui 25.906.906 linhas com 1.2 GB de armazenamento físico, por razões de processamento foi utilizado um arquivo secundário com o subconjunto de dados sendo este arquivo com 909,209 linhas com 55.636 MB.

Conforme ilustrado na Figura 6, o *dataset* apresentado nesse artigo está em evolução, serão agregados outros dados como os conceitos para cada termo e classificações e conceitos para os termos. Outro ponto que será evoluído no *dataset* refere-se às *Uniform Resource Identifier* ou *URIs* associadas a cada um dos termos coletados. As *URIs* são os identificadores únicos dos recursos na *Web* e facilitam o mapeamento direto entre os termos do subconjunto proposto com as demais informações representadas no *AGROVOC*, esses dados serão agregados à tabela de termos.

A Figura 6 ainda apresenta uma ilustração de algumas classificações de termos em idioma original (inglês) mostrando as hierarquias e um exemplo de conceito associado a um dos termos. Apesar de já existirem alguns mecanismos para lidar diretamente com arquivos *RDF* para realizar o tratamento e obtenção de dados customizados como bibliotecas específicas *RDFLib - Python* ou consultas *SPARQL*, optou-se por desenvolver uma implementação própria de forma simples sem a exigência de recursos de terceiros,

**Figura 6. Modelo ExR Resultante**




da mesma forma o resultado final é um *dataset* simples facilmente acoplado a qualquer base de dados ou acessível com interfaces simples de consulta.

## 6. Conclusões e Trabalhos futuros

Parte do processo de criação científica consiste em consultas a termos técnicos específicos que normalmente não segue um vocabulário comum ou traduções diretas entre idiomas. Problemas relacionados a erros de tradução de termos técnicos são frequentemente observados quando realizados por profissionais sem domínio técnico específico, dessa forma, o resultado pode não ser satisfatório. A estrutura do *dataset* descrito é simples, não havendo necessidade de grandes requisitos para uso em qualquer aplicação, sendo necessário apenas o *driver* para estabelecer a conexão entre dados e aplicação. Conforme descrito no artigo, há interfaces para consulta à ontologia *Agrovoc*, entretanto a ferramenta a simplificação dessa consulta em uma estrutura tabular com poucas colunas proporciona agilidade tanto para a criação de consultas rápidas quanto para eventuais integrações com ferramentas de terceiros como *Enterprise Resource Planning* ou ferramentas para gestão de informações como *business intelligence*. Uma oportunidade proposta para trabalhos futuros é a criação de uma interface gráfica para consultas a partir de qualquer documento com armazenamento de termos local, nesse sentido, o dicionário de termos estará sempre disponível para o usuário final. Nesse contexto, o impacto direto dessa proposta é a contribuição para profissionais, acadêmicos e de áreas afins.

Outro ponto importante refere-se aos dados disponíveis no *dataset*, inicialmente foram criados para uso local, não foi aventado a possibilidade de uso de forma distribuída ou utilizando algum tipo de publicação em ambiente Web como *webservice*. Entretanto, é importante destacar que a proposta apresentada neste artigo é uma versão inicial do *dataset* e não exclui ou esgota as possibilidades de uso de forma distribuída. A versão inicial do projeto é a criação do *dataset* podendo evoluir para extensões complementares como aplicações para acesso aos dados.

## Referências

AgXML (2022). Agxml - agriculture information xml schema documentation. `https://schemas.liquid-technologies.com/agxml/2.0/`. (Accessed on 07/07/2022).

Arp, R., Smith, B., and Spear, A. D. (2015). Building ontologies with basic formal ontology. *The MIT Press*.

Aryal, J., Morshed, A., and Dutta, R. (2014). Extracting urban forests in geobia framework using agrovoc and worldview-2 imagery. In *GEOBIA*.

Beneventano, D., Bergamaschi, S., Sorrentino, S., Vincini, M., and Benedetti, F. (2013). Semantic annotation of the cerealab database by the agrovoc linked dataset. In *Ecol. Informatics*.

Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., and Keizer, J. (2013). The agrovoc linked dataset. *Semantic Web*, 4:341–348.

Dermeval, D., Vilela, J., Bittencourt, I. I., de Castro, J. B., Isotani, S., da S. Brito, P. H., and Silva, A. (2015). Applications of ontologies in requirements engineering: a systematic review of the literature. *Requirements Engineering*, 21:405–437.

eng Fao, R. . A. S. S. D. (2003). Agricultural metadata element set (agmes). namespace documentation for document-like information objects. *FAO*.

Maron, D. and Feinberg, M. (2018). What does it mean to adopt a metadata standard? a case study of omeka and the dublin core. *J. Documentation*, 74:674–691.

Martini, D., Schmitz, M., and Mietzsch, E. (2013). agrordf as a semantic overlay to agroxml : a general model for enhancing interoperability in agrifood data standards. *CIGR International Commission of Agricultural and Biosystems Engineering*.

Mietzsch, E., Martini, D., Kolshus, K., Turbati, A., and Subirats, I. (2021). How agricultural digital innovation can benefit from semantics: The case of the agrovoc multilingual thesaurus. *Engineering Proceedings*.

NISO (2022). National information standards organization — niso website. `https://www.niso.org/`. (Accessed on 07/06/2022).

Salokhe, G., Onyancha, I., Weinheimer, J., Ward, F. L. H., and Keizer, J. (2005). The agris application profile for the international information system on agricultural sciences and technology. *FAO*.

Simek, P., Vanek, J., Jarolimek, J., Stoces, M., and Vogeltanzova, T. (2018). Using metadata formats and agrovoc thesaurus for data description in the agrarian sector. *Plant Soil and Environment*, 59:378–384.

Wieczorek, J., Bloom, D., Guralnick, R. P., Blum, S. D., Döring, M., Giovanni, R. D., Robertson, T., and Vieglais, D. (2012). Darwin core: An evolving community-developed biodiversity data standard. *PLoS ONE*, 7.