

Correção: Campo 'Published Year'

01. DADOS GERAIS

Linguagem de Programação:	Scala
Build:	SBT (Scala Build Tool)
Categoria:	Correção

02. ERRO/DEFEITO

1. Chave 'PublishedYear' dos documentos gerados no mongoDB pelo projeto SparkCovid está com as palavras unidas conforme imagem abaixo:

```
_id: ObjectId('646b566026d03c059f386030')
Abstract: "SARS-CoV-2-mediated interactions with drug metabolizing enzymes and me..."
Accession Number: "37180730"
AlternateId: ""
Authors: "Nwabufu, C. K. | Hoque, M. T. | Yip, L. | Khara, M. | Mubareka, S. | P..."
CovNum: "2317756.0"
Database: "MEDLINE"
Date Added: ""
DOI: "10.3389/fphar.2023.1124693"
FulltextLink: "https://doi.org/10.3389/fphar.2023.1124693"
Issue: ""
Journal: "Frontiers in Pharmacology"
Keywords: ""
KJD: ""
Language: "en"
Pages: "1124693"
PMID: "37180730"
PublishDate: ""
Published Month: ""
PublishedYear: "2023"
SCIELO: ""
Tags: ""
Title: "SARS-CoV-2 infection dysregulates the expression of clinically relevan..."
UNKNOWN: ""
Volume: "14"
WOS: ""
_updd: 2023-05-22T03:00:00.000+00:00
_upddSrc: 2023-05-19T03:00:00.000+00:00
```

2. A incorreta definição da chave foi encontrada através do módulo 'ScalaMongoDiff' comparando duas coleções no mongoDB. Uma coleção incluída no mongoDB pelo SparkCovid e outra pelo MongoDBMigrations lendo diretamente das planilhas do PENTAHO.

```
_id: ObjectId('6463b140bfa32145e52f03be')
CovNum: "1565083"
Date Added: Array
  0: ""
  1: "16/09/2022"
Published Year: Array ← Correto
  0: ""
  1: "2021"
PublishedYear: Array ← Incorreto
  0: "2021"
  1: ""
_updd: "Tue May 16 13:37:20 BRT 2023"
```

03. ESPERADO

1. É esperado a chave 'PublishedYear' com a separação das palavras Published e Year conforme a imagem abaixo:

```
_id: ObjectId('646b566026d03c059f386030')
Abstract: "SARS-CoV-2-mediated interactions with drug metabolizing enzymes and me..."
Accession Number: "37180730"
AlternateId: ""
Authors: "Nwabufo, C. K. | Hoque, M. T. | Yip, L. | Khara, M. | Mubareka, S. | P..."
CovNum: "2317756.0"
Database: "MEDLINE"
Date Added: ""
DOI: "10.3389/fphar.2023.1124693"
FulltextLink: "https://doi.org/10.3389/fphar.2023.1124693"
Issue: ""
Journal: "Frontiers in Pharmacology"
Keywords: ""
KJD: ""
Language: "en"
Pages: "1124693"
PMID: "37180730"
PublishDate: ""
Published Month: ""
Published Year: "2023"
SCIELO: ""
Tags: ""
Title: "SARS-CoV-2 infection dysregulates the expression of clinically relevan..."
UNKNOWN: ""
Volume: "14"
WOS: ""
_updd: 2023-05-22T03:00:00.000+00:00
_upddSrc: 2023-05-19T03:00:00.000+00:00
```

04. CORREÇÃO NO CÓDIGO

1. No objeto SparkCovidWho e metodo main deve se corrigir os nomes das chaves para 'Published Year'.

```
SparkCovidWho.scala x
36
37 val df1: DataFrame =
38   df.withColumn("Abstract", Abstract._udf(col("Abstract")))
39   .withColumn("Accession number", AccessionNumber._udf(col("Accession number")))
40   .withColumn("Author", Authors._udf(col("Author"))).withColumnRenamed("Author", "Authors")
41   .withColumn("Refid", Abstract._udf(col("Refid"))).withColumnRenamed("Refid", "CovNum")
42   .withColumn("Database", Database._udf(col("Database")))
43   .withColumn("Doi", DOI._udf(col("Doi")))
44   .withColumn("Issue", Issue._udf(col("Issue")))
45   .withColumn("Journal", Journal._udf(col("Journal")))
46   .withColumn("Keywords", Keywords._udf(col("Keywords")))
47   .withColumn("Language", Language._udf(col("Language")))
48   .withColumn("Pages", Pages._udf(col("Pages")))
49   .withColumn("Published Month", PublishedMonth._udf(col("Publishdate")))
50   .withColumn("PublishedYear", PublishedYear._udf(col("Year")))
51   .withColumn("Title", Title._udf(col("Title")))
52   .withColumn("Volume", Volume._udf(col("Volume")))
53   .withColumn("_upddSrc", UpddSrc._udf(col("_updd")))
54   .withColumn("_updd", lit(date))
55   .withColumn("PMID", PMID._udf(col("Accession number")))
56   .withColumn("WOS", WOS._udf(col("Accession number")))
57   .withColumn("KJD", KJD._udf(col("Accession number")))
58   .withColumn("SCIELO", SCIELO._udf(col("Accession number")))
59   .withColumn("UNKNOWN", UNKNOWN._udf(col("Accession number")))
60   df1.printSchema()
61
62 val df2: DataFrame =
63   df1.withColumn("FulltextLink", FulltextLink._udf(col("Doi"), col("PMID")))
64   df2.printSchema()
65
66 val df3: DataFrame = df2.select("Abstract", "Accession Number", "AlternateId", "Authors", "CovNum",
67   "Database", "Date Added", "DOI", "Fulltextlink", "Issue", "Journal", "Keywords", "KJD", "Language", "Pages", "PMID",
68   "PublishDate", "Published Month", "PublishedYear", "SCIELO", "Tags", "Title", "UNKNOWN", "Volume", "WOS", "_updd",
69   "_upddSrc")
70   df3.printSchema()
71   df3.show(3, truncate = true)
72
```