CHALLENGES IN CLOUD INFRASTRUCTURE AND SCIENTIFIC DATA: TECHNICAL PROPOSALS AND TOOLS APPLIED TO THE SCIELO DATABASE

Desafios em infraestrutura de nuvem e dados científicos: propostas técnicas e ferramentas aplicada para a base de dados SciELO

Desafíos en infraestructura en la nube y datos científicos: propuestas técnicas y herramientas aplicadas a la base de datos SciELO

Alysson Fernandes Mazoni
Doctor, University of Campinas (Unicamp), Campinas, São Paulo, Brazil.
Post-doctoral researcher.
Lattes: http://lattes.cnpq.br/0347768173274366
ORCID: https://orcid.org/0000-0001-5265-6894

João de Melo Maricato
Doctor, University of Brasilia (UnB), Brasília, Distrito Federal, Brazil.
Professor.
Lattes: http://lattes.cnpq.br/3991129099537472
ORCID: https://orcid.org/0000-0001-9162-6866

Ronaldo Ferreira de Araújo
Doctor, Federal University of Alagoas (UFAL), Maceió, Alagoas, Brazil.
Professor.
Lattes: http://lattes.cnpq.br/3328212638040851
ORCID: https://orcid.org/0000-0003-0778-9561

Rodrigo Costas Comesaña
Doctor, Centre for Science and Technology Studies (CWTS), University of Leiden, Leiden, South Holland, The Netherlands.
Senior researcher.
Stellenbosch University, Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (DSI-NRF-SciSTIP), Stellenbosch, South Africa.
Extraordinary professor.
Lattes: http://lattes.cnpq.br/8677394986541681
ORCID: https://orcid.org/0000-0002-7465-6462

Resumo

Introdução: Aplicações de computação em nuvem tem relevância significativa para bases de dados científicas baseadas em dados de pesquisa. Este trabalho descreve os passos e requisitos de aplicação de tais tecnologias para a base de dados do SciELO. **Metodologia**: Utilizamos infraestrutura comercial em nuvem e propusemos

uma sequência de passos para construir um modelo de dados que a partir de bases de dados não estruturadas. **Resultados**: Investigamos iterativamente os dados da base de dados SciELO e construímos um modelo relacional. Todos os códigos estão disponíveis publicamente. **Conclusão**: A despeito das preocupações com a soberania de dados, serviços comerciais de computação em nuvem são uma boa opção para projetos de curto prazo. Além disso, a sequência de etapas e a estrutura geral dos códigos aqui desenvolvidos podem ser usados para tornar outras bases de dados semelhantes disponíveis e úteis.

Palavras-chave: análise de dados; ciclo de vida dos dados; recuperação da informação.

Abstract

***Introduction**: Cloud computing applications have significant relevance for scientific databases based on research data. This work aims to report the proposed requisites and steps of technologies and tools to develop cloud computing applications based on research data for the SciELO database. **Methods**: We have used commercial cloud infrastructure and proposed a step-by-step method that could be applied to unstructured databases aiming at a comprehensive data model. **Results**: We have iteratively looked at data entries and developed SQL scripts that build a data model for the SciELO database. All the scripts are made public. **Conclusion**: Despite the data sovereignty concerns, commercial cloud services are a good option for short term projects. Also, the sequence of steps and general structure of the scripts can be used to make other similar open databases available and useful.*

***Keywords**: data analytics; data lifecycle; information retrieval.*

Resumen

***Introducción**: Las aplicaciones de computación en la nube tienen una relevancia significativa para las bases de datos científicas basadas en datos de investigación. Este trabajo describe los pasos y requisitos para aplicar dichas tecnologías a la base de datos SciELO. **Metodología**: Utilizamos infraestructura de nube comercial y propusimos una secuencia de pasos para construir un modelo de datos basado en bases de datos no estructuradas. **Resultados**: Investigamos iterativamente datos de la base de datos SciELO y construimos un modelo relacional. Todos los códigos están disponibles públicamente. **Conclusión**: a pesar de las preocupaciones sobre la soberanía de los datos, los servicios comerciales de computación en la nube son una buena opción para proyectos a corto plazo. Además, la secuencia de pasos y la estructura general de los códigos desarrollados aquí pueden usarse para hacer que otras bases de datos similares estén disponibles y sean útiles.*

***Palabras clave**: análisis de datos; ciclo de vida de datos; recuperación de información.*

DOI 10.22477/vi.widat.44

1. Introduction

Data resulting from scientific activities, present in bibliographic databases, are widely used to produce scientometric and altmetric indicators and information retrieval. These types of data are, to a large extent, semi-structured and heterogeneous, therefore, more complex to use when compared to structured data (BOUGANIM; GALHARDAS; MANOLESCU, 2021). A systematic problem of bibliographic databases is the fact that most data are inserted manually, creating potential inconsistencies and errors. Another important point is that definitions of patterns for data storage and manipulation are just recently becoming known in many fields of application. Due to this fact, many important sources of data, even open data, are difficult to use given their file formats, inconsistencies in indexing and demand for high processing power.

Classical scientometric databases like Web of Science (WoS) or Scopus have historically received numerous criticisms. The lack of comprehensive coverage is the most common and important criticism (MELO; TRINCA; MARICATO, 2021; SIMONS, 2008; SANTOS *et al.*, 2021). This is one of the reasons for creating specific indexes and databases in Latin America as well as other peripheral countries (SANTIN; CAREGNATO, 2018).

In this context, in 1998, the Scientific Electronic Library Online (SciELO) was created. It emerged as a program to support the Brazilian research infrastructure, which selects journals published by the country's institutions. The SciELO database has unique analytical characteristics that need to be valued. It is a source for open access scientific publication in the regions that it covers. SciELO is relevant, as it contains metadata elements that may not exist in other databases, like for example, acknowledgments, counts on views and downloads, submission to publication dates, etc. (MARICATO *et al.*, 2023). According to the library's website, currently (April 2023), 1,905 journals, more than 1 million documents and 27.5 million references are indexed, from which, SciELO Brazil contributes with 399 journals, 483,541 documents and 12,749,068 references.

The SciELO database, although using metadata such as XML and JSON, does not yet have fully structured data and, consequently, does not have a robust API available. In bibliographic databases, in many cases, the definition of minimum requirements and steps is necessary for the available data to become useful.

This work aims to report the proposed requisites and steps of technologies and tools to develop of cloud computing applications based on research data from SciELO data. Thus, in the present research, considering the characteristics of the heterogeneous data of the SciELO database, we understand that the proposal for the construction of a data infrastructure should consider requirements and steps:

- Computing infrastructure, both in software and hardware, with capacity to store and process a large amount of data, with possibility of performing queries using SQL.

- Technological capabilities to read and store unstructured data (HTML, XML and JSON formats, mostly) in a way that their sorting and finding is within available computational limits. This step demands computational infrastructure both in software and hardware.

- Possibilities to map and convert the data into a relational or tabular model. This is the step that structures the data and allows descriptive analysis.

- Make the structured data available for further work, preferably over the internet in an open platform.

The first three aspects require a computational infrastructure able to store the data and keep it in memory to parse them using algorithms. Given the size, this can be a limitation to be overcome with large memory or even distributed infrastructure such as is possible with the Hadoop collection of software packages (LAM, 2010). The third aspect involves a database server, usually compatible with SQL, a web services provider and an API to connect both, building an interface. Depending on the scale, this part can be deployed using a container orchestration system such as Kubernetes (LUKSA, 2017).

2. Methods

In our study of the SciELO data[1], we wanted to avoid infrastructure work, we have chosen to use Google Cloud Platform and some of its tools (BISONG, 2019a). Using a commercial platform gives access to storage space, virtual machines configured as needed and a SQL based warehouse such as BigQuery. One of the institutions involved (University of Campinas) subscribes to Google Services, making it a simpler option.

In order to download, read and convert the original files, we have used the Python distribution as installed inside the *Colaboratory*. Google *Colaboratory* is a literate programming interface based on Jupyter Notebok and the Anaconda distribution of Python packages (CARNEIRO *et al.*, 2018).

Considering the format of the data provided by SciELO, we made some choices regarding its conversion. Along the time, the XML was proposed to separate the data from the visual structure of a file. Along the time, the XML was proposed to separate the data from the visual structure of a file. We used, in this research, XML, HTML and JSON formats. They are to be able to convert information in a hierarchical organization. JSON format was proposed in the context of Javascript language that is the most common to this day when developing WWW. Other tools and languages have adopted some form of library to write JSON files from raw data as well as converting to and from other formats such as HTML and XML.

Once the files are converted into JSON, it is possible to generate a simple table with JSON formatted text fields. Each one of the files becomes a row in a table with a single column. That table is then uploaded onto the Google BigQuery platform (BISONG, 2019b). From that point, it is possible to use SQL JSON functions (PETKOVIĆ, 2017; LIU, 2019; ZHANG *et al.*, 2022) that manage to collect fields from JSON entries and extract them as column values. Tools for such tasks are available in most SQL database warehousing services given the fact that unstructured data is most frequent (MARRS, 2017).

Our method for creating a data model from the SciELO data can be summarized as follows:

I. Collect portion of raw files;

II. List important fields identified;

III. Map fields that should be in different tables with an index to the main table;

IV. Write or update scripts to assemble tables for the portion of files;

1 The bulk of SciELO works as an XML file for each entry is available through a public link: http://static.scielo.org/articlemeta/scielo_articles.zip. Acesso em: 27 jun. 2023.

V. Add extra portion of raw files and repeat the process.

The incremental details of the data model are constantly judged for their completion until the point that all relevant metadata is considered sufficient.

3. Results and discussion

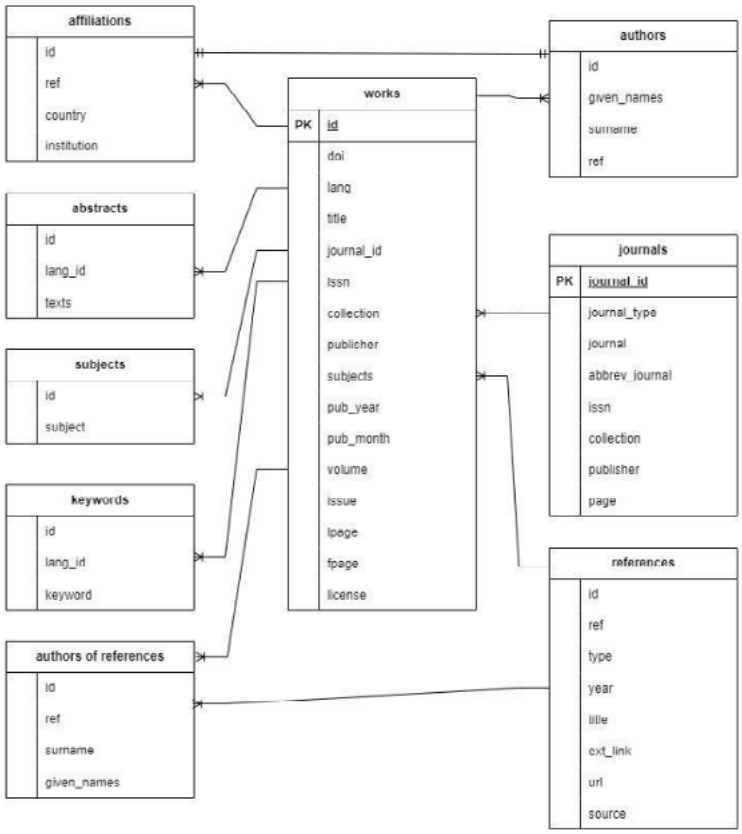
A difficulty that usually arises when dealing with unstructured data is the fact that field names are not always consistent. For example, the name used for the title of works and its translated form are sometimes stored in different field names inside the SciELO metadata dump. This is a result of changes in the way the original files were stored (firstly HTML, then XML, now a non-SQL database entry) and the different sources of the metadata.

That forces us to count how many works are collected with incomplete information and then, using the identifier, look back in the JSON entry to locate the exact name that is missing. When the field name is found, it is added to the script and all variants are coalesced into a single column. It is also common that some fields are of variable number of pieces of information, that is usually an indication that they should belong to another table linked to the current by an identifier. However, as an intermediate step, one can choose to keep the field in the JSON format in a column allowing for later generation of another table or further querying using JSON functions if the data present is still unstructured. This was done here in the keywords field (which is also a different table).

Since our aim is to produce a data model, we have to look through examples of XML files and list their fields. By looking at other databases, such as OpenAlex (PRIEM; PIWOWAR; ORR, 2022), we have similar patterns to follow. So, there must be a table listing the works and their most direct information such as publication date and title. Eventually, we understand that other fields are linked to the work but have other information associated with them that should be stored in other tables and linked back to the table of works using an identifier. Examples such as this are authors, affiliations and references.

The tables are connected by the identifiers described and can be seen as a data model in Figure 1.

Figure 1 - Proposal for the SciELO data model.



Source: Research data (2023).

The scripts with all the steps are made public in a Github[2] repository. Regarding the update policy, it can be determined by the user, in case he uses the codes. In this case, if the steps are repeated starting with the download from SciELO, tables in the most recent version of the data are produced. For researchers interested in using the public tables as provided by the authors in Google BigQuery, they are updated every two months or according to the authors needs, and the date of the available version can be seen as the last update in the tables presented.

When collecting possible fields in the table, as the code in table 4, we got 528,976,138 fields with an average of 498.22 fields each entry. We have achieved the list of tables and their columns expressed in Table 1.

The tables in the BigQuery platform as described in Table 1 are on-line and available for research. Also, the scripts allow regeneration of the tables from the original data. As can be seen, SciELO in the version of the dump that we used contains 1,061,736 works and 3,211,988 authorships, making and average of 3,2 authors each work.

2 Disponível em: https://github.com/alyssonmazoni/scielo_relational_model. Acesso em: 27 jun. 2023

Table 1 - Tables extracted from the SciELO data.

TABLE/SIZE	COLUMNS	DESCRIPTION
works 1,061,736	*id, doi, title, lang, type, *journal_id, journal_type, journal, abbrev_journal, issn, collection, publisher, subjects, pub_year, pub_month, volume, issue, lpage, fpage, license, keywords	Main table with information of the work and identifier.
authors 3,211,988	*id, surname, given_names, *ref	Authorships: authors as linked to the works. Work is referenced by *id*. Affiliation is referenced by *ref*.
affiliations 1,703,596	*id, *ref, institution, country	Affiliations, *id* references the works and *ref* links the author.
abstracts 847,589	*id, lang_id, text	Abstracts of works, it is possible to have more than one given different languages.
subjects 979,858	*id, subject	Classification of subjects proposed by SciELO, *id* also links the work.
keywords 6,517,337	*id, lang_id, keyword	Keywords provided by authors or publishers. *id* links the work.

Source: Research data (2023).

4. Conclusions

Overall, cloud computing applications provide the necessary infrastructure, scalability, cost efficiency, accessibility, and advanced analytics capabilities to enhance scientific databases' management, analysis, collaboration, and long-term preservation of research data (LANEY, 2001). In this work, we have used cloud commercial platforms as tools to create a structured model for the SciELO data, called Google Cloud Platform.

We have managed to keep data format and scripts open and written in commonly used languages (Python and SQL), allowing a possible implementation on different platforms as well as on local servers using similar open-source tools.

Another important benefit of this work is having a set of scripts and a public link to access SciELO metadata in a structured fashion. Large scale bibliometric research can be done with this tool. The fact that the database is also on a major cloud service provider, despite the mentioned problems, contributes to its availability to broader communities of interested users. It also allows APIs and other access interfaces to be developed over the database. The same steps are general for other databases with the same character of unstructured data, it being possible to adapt them.

The developed solutions could be used by the SciELO database, with a view to providing its open data through an API.

Further investigations and future analyses deserve to be carried out, since the data are now available with free and relatively easy access. The quality, scope, completeness, level of disambiguation and standardization of data extracted from SciELO databases are relevant research objects that can be carried out in the future. Even so, the data can already be used in several researches, considering that the currently available databases represent an alternative for the analysis of science, especially in peripheral countries.

References

BISONG, Ekaba. An overview of Google Cloud Platform Services. *In:* BISONG, Ekaba. **Building machine learning and deep learning models on Google Cloud Platform**. Berkeley: Apress, 2019. p. 7-10. Disponível em: http://dx.doi.org/10.1007/978-1-4842-4470-8_2. Acesso em: 27 jun. 2023.

BISONG, Ekaba. Google BigQuery. *In:* BISONG, Ekaba. **Building machine learning and deep learning models on Google Cloud Platform**. Berkeley: Apress, 2019. p. 485-517. Disponível em: http://dx.doi.org/10.1007/978-1-4842-4470-8_38. Acesso em: 27 jun. 2023.

BOUGANIM, Théo; GALHARDAS, Helena; MANOLESCU, Ioana. Efficiently identifying disguised nulls in heterogeneous text data. *In:* BDA (CONFÉRENCE SUR LA GESTION DE DONNÉES - PRINCIPES, TECHNOLOGIES ET APPLICATIONS), 37., 25-28 Oct. 2021, Paris. **Communication** [...]. Paris: [*s. n.*], 2021. Disponível em: https://inria.hal.science/hal-03347947. Acesso em: 27 jun. 2023.

CARNEIRO, Tiago *et al.* Performance analysis of google colaboratory as a tool for accelerating deep learning applications. **IEEE Access**, [Piscataway], v. 6, p. 61677-61685, 2018. Disponível em: https://doi.org/10.1109/ACCESS.2018.2874767. Acesso em: 6 jul. 2023.

LAM, Chuck. **Hadoop in action**. [*S. l.*]: Manning, 2010. 325 p.

LIU, Zhen Hua. JSON Data Management in RDBMS. *In:* MA, Zongmin; YAN, Li (ed.). **Emerging technologies and applications in data processing and management**. Hershey, PA: IGI Global, 2019. p. 20-44. Disponível em: https://doi.org/10.4018/978-1-5225-8446-9.ch002. Acesso em: 6 jul. 2023.

LUKSA, Marko. **Kubernetes in action**. [S. l.]: Manning, 2017. 324 p.

MARICATO, João de Melo *et al.* SciELO as an open scientometric research infrastructure: general discussion of coverage in OpenAlex, WoS, Scopus and Dimensions [preprint]. 27TH INTERNATIONAL CONFERENCE ON SCIENCE, Technology and Innovation Indicators (STI 2023).

MARRS, Tom. **JSON at work**: practical data integration for the web. Sebastopol, CA: O'Reilly Media, 2017. 376 p.

MELO, João Henrick Neri de; TRINCA, Tatiane Pacanaro; MARICATO, João de Melo. Limites dos indicadores bibliométricos de bases de dados internacionais para avaliação da pós-graduação brasileira: a cobertura da Web of Science nas diferentes áreas do conhecimento. **Transinformação**, Campinas, v. 33, e200071, 2021. Disponível em: https://doi.org/10.1590/2318-0889202133e200071. Acesso em: 6 jul. 2023.

PETKOVIĆ, Dušan. JSON integration in relational database systems. **International Journal of Computer Applications**, [Bangalore], v. 168, n. 5, p. 14-19, June 2017. Disponível em: https://doi.org/10.5120/ijca2017914389. Acesso em: 6 jul. 2023.

PRIEM, Jason; PIWOWAR, Heather; ORR, Richard. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. *In:* INTERNATIONAL CONFERENCE ON SCIENCE AND TECHNOLOGY INDICATORS, 26., 7-8 September 2022, Granada, Spain. **STI Conference** [...]. Spain, 2022. Disponível em: https://doi.org/10.48550/arXiv.2205.01833. Acesso em: 27 jun. 2023.

SANTIN, Dirce Maria; CAREGNATO, Sônia Elisa. Índices de citação nacionais e regionais: importância, experiências e perspectivas para a América Latina. *In:* ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 6., 17-20 jul. 2018, Rio de Janeiro. **Anais [...]**. Rio de Janeiro: UFRJ, 2018. p. 54-62. Disponível em: https://www.lume.ufrgs.br/handle/10183/183984. Acesso em: 27 jun. 2023.

SANTOS, Solange Maria dos *et al.* The relationship between the publication language and its impact on public and collective health. version 1. **SciELO Preprints**, 2020. Disponível em: https://doi.org/10.1590/SciELOPreprints.1549. Acesso em: 6 jul. 2023.

SIMONS, Kai. The misused impact factor. **Science**, Washington, v. 322, n. 5899, p. 165, 10 Oct. 2008. Disponível em: https://dx.doi.org/ 10.1126/science.1165316. Acesso em: 6 jul. 2023.

ZHANG, Lei *et al.* JSON-based control model for SQL and NoSQL data conversion in hybrid cloud database. **Journal of Cloud Computing**, [Heidelberg], v. 11, article number 23, 2022. Disponível em: http://dx.doi.org/10.1186/s13677-022-00302-9. Acesso em: 27 jun. 2023.