alter ibi

n.1, 2014 / ISSN 2183-2927 CIEBA-FBAUL / UTAD

Behind a 'NoSQL' approach in the development of a bibliography 'captabase' for rupestrian imagery

Ludwig Jaffe
Mila Simões de Abreu
Cris Buco
Maxim Jaffe

Projeto 4 Dimensões, Coronel José Dias, Piauí (Brasil) Universidade Trás-os-Montes e Alto Douro (UTAD), Unidade de Arqueologia, (Portugal) Centro de Estudos Transdisciplinares para o Desenvolvimento (CETRAD), Vila Real (Portugal)

E-mail: project4dimension@gmail.com

Resumo: Este artigo revela os principais componentes que estão por trás de uma aplicação web para um chamado 'captabase' bibliográfico, bem como fornecer informações sobre os termos técnicos e outras palavras menos comuns.

Palavras chave: 'NoSQL' bibliografia 'capta' 'captabase' rupestre imagens

Abstract: This paper discloses key components that lie behind a web application for a bibliographic 'captabase', as well as providing information about strange terms and other unusual words.

Keywords: 'NoSQL' bibliography 'capta' 'captabase' rupestrian imagery

Data Base

The I Congress on "Global Heritage Management" (24–25 February 2014, Auditório do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, São Raimundo Nonato, Piauí, Brazil), provided an opportunity to present a web application entitled "Bibliografia da Serra da Capivara". This paper discloses an array of constituent components that lie behind this bibliography 'captabase', as well as providing information about strange terms and other unusual words.

Words and meaning

> 'When I use a word,' Humpty Dumpty said, in rather a scornful tone, 'it means just what I choose it to mean — neither more nor less.' — Lewis Carroll, *Through the Looking Glass, and What Alice Found There* (1872) (CARROLL 1917, 99)

What is NoSQL?

SQL is an acronym for Structured Query Language (GENNICK 2010), a programming language for managing information held in a 'Relational Database Management System' (RDBMS) (CODD 1970). In 1998, Carlo Strozzi coined the somewhat provocative term, NoSQL (HAUGEN 2010). Over a decade later, in June 2009, Johan Oskarsson and Eric Evans reused the NoSQL term (EVANS 2009, ROE 2012) to refer to the rapidly multiplying number of distributed, non-relational systems that generally do not conform to the four features of traditional relational database systems: atomicity, consistency, isolation, durability (ACID) (CHAPPLE 2014). Strozzi quips that the new NoSQL movement should have been called 'NoREL' (STROZZI 2010).

About capta

> "Not data but capta."
> — Christopher Chippindale

The word 'capta' was possibly first coined in 1991 when Christopher Chippindale argued the correctness of that word in an editorial of *Antiquity* (CHIPPINDALE 1991, 442).

The word was adopted or re-coined by Peter Checkland and Sue Holwell in the book, *Information, Systems and Information Systems: Making Sense of the Field* (CHECKLAND and HOLWELL 1997). For Checkland and Holwell, data represents all those masses of facts, observations and concepts that exist in the universe. Once data are captured as part of an information system, a conversation or any kind of interaction, they become capta.

In 2000, Chippindale returns to the the defense of the term in an article of *American Antiquity*, Capta and Data: On the True Nature of Archaeological Information (CHIPPINDALE 2000):

> "(…) they are practically never given to us by the archaeological record. They are actually capta, things that we have ventured forth in search of and captured (…)

Rupestrian advocacy

In 1989, members of the Cooperativa Archeologica «Le Orme dell'Uomo» advocated the term "rupestrian archaeology" to describe an archaeological approach to such material (FOSSATI 1997).

Dictionary definitions show this use of the word, rupestrian, to be appropriate (Rupestrian 2014):

> (Of art) done on rock or cave walls.
> late 18th century: from modern Latin *rupestris* 'found on rocks' (from Latin *rupes* 'rock')

Why imagery?

Why the word, imagery? Dialog with indigenous people in North America found that they were uncomfortable with the word "art" as a description for the creations of their ancestors, preferring the word imagery (DEAN 2006, 10–11).

> *The use of the word art is something I have a particular beef about, and this comes directly from the folks that I work with. The tribal elders in the region where I live in the Pacific Northwest asked me not to use that term, because they find it offensive. That is the case elsewhere, although dislike of the term is not universal. Personally, I am uncomfortable with the word art for this type of work and use the term imagery instead, partly in deference to my elders but also because my work has taught me that there is something else here.*— J. Claire Dean

Revista ALTER IBI. ISSN 2183-2927. Vol. 1 (1): 143-150.

Tables and trees

Established relational models have rigid table-like structures. So-called 'document-oriented' frameworks tend to be tree-like or hierarchical and better suited to semi-structured information. Consequently, methods for querying and retrieving information from such systems differ from those requiring a relational approach.

XML and JSON

Two widely adopted standard formats or encodings characterize semi-structured information: XML (Extensible Markup Language) (HAROLD and MEANS 2004) and JSON (JavaScript Object Notation) (CROCKFORD 2008).

```
<personname>
  <firstname>Aziz</firstname>
  <surname>Ab'Saber</surname>
</personname>
```

A fragment of XML

```
{
 "personname": {
  "firstname": "Aziz",
  "surname": "Ab'Saber"
 }
}
```

A fragment of JSON

Before XML and JSON

Long before XML, at IBM (International Business Machines), Charles Goldfarb, Edward Mosher and Raymond Lorie developed GML (Generalized Markup Language), an acronym Goldfarb invented by using their surname initials (ANDERSON 2001). Goldfarb acknowledges the credit due to William Tunnicliffe's inspiration presentation about the separation of information content of documents from their format (GOLDFARB 1997). A descendant of GML emerged, a complex specification called SGML (Standard Generalized Markup Language). SGML was the starting point for a simpler implementation, HTML (HyperText Markup Language). In the summer of 1991, while working at CERN (formerly, Conseil Européen pour la Recherche Nucléaire), Tim Berners-Lee released the first web server, which disseminated specifications of HTML (BERNERS-LEE 1998).

Why XML

Developers soon found the much welcomed simplicity of HTML rather limiting, so leading to XML and XHTML (Extensible HyperText Markup Language) (ROBBINS 2010). Like SGML, XML allows for the separation of presentation and content. Besides its suitability for automated processing, XML is relatively readable by humans when neatly indented. This is can be vital when automated or proprietary systems fail to retrieve required information. As always, whatever systems are in use, it is prudent to make backups, saving or exporting the capta as XML formatted text.

Querying XML

Just as SQL was designed for relational structures, SGML had DSSSL (Document Style Semantics and Specification Language) and XML has several: XPath (XML Path Language) and XPointer (XML Pointer Language) (SIMPSON 2002), XSLT (Extensible Stylesheet Language Transformations) (TIDWELL 2008) and XQuery (WALMSLEY 2007).

XSLT and XForms

The authors presented proof of concept applications deploying XSLT and XForms (XML Forms) at meetings of the British Rock Art Group, one talk entitled "Databases Without Databases for the World of Rock Art" (JAFFE 2007), the other, "Captabases without Captabases for Rupestrian Imagery" (JAFFE 2010). XForms is another member of the XML family that has methods to parse XML documents, update chosen elements and insert new ones (DUBINKO 2003).

XQuery

The XQuery language is designed so that queries are concise and easily understood; the language includes methods for updating XML documents and inserting new nodes. Implementations such as eXist-db (EXIST-DB 2014) provide further tools to create new documents and the integrated web server delivers the means to present results of the queries.

Revista ALTER IBI. ISSN 2183-2927. Vol. 1 (1): 143-150.

DocBook

There are numerous XML schemes. Rather than reinventing the wheel with yet another markup arrangement, efforts focused on finding an established well-documented project that best suited the development aims. DocBook was a good match (WALSH 2010) as it has a plethora of open source tools to transform DocBook XML to other formats such as PDF and HTML using the tools like the Velocity DocBook framework (D'ABREO 2007).

Bibliography of the Serra da Capivara, Piauí, Brazil

Studies on a 'NoSQL' approach in the development of a bibliography 'captabase' for rupestrian imagery has been quietly advancing for several years, but last year an opportunity arose to create a working implementation. The project tested two of the most promising possibilities: BaseX (BASEX 2014) and eXist-db . Despite its features and compact distribution size, BaseX seemed unstable, so eXist-db became the development platform.

Initial content consisted of a bibliography written in a LibreOffice ODT file compiled by two researchers, Cristiane Andrade Buco and Mila Simões de Abreu. Entries were formatted in a slightly modified Chicago Manual of Style (CMOS) author-date style (UNIVERSITY OF CHICAGO PRESS STAFF 2010). There was much duplicated information, so Maxim Simões de Abreu Jaffe cut this to an essential list of author(s)-date entries and used the regular expressions (STUBBLEBINE 2007) integrated in search and replace function of LibreOffice to semi-automatically transform entries to DocBook XML. This he saved as plain text and transformed it using an XSLT (Extensible Stylesheet Language Transformations) style sheet to 799 separate files, ready for upload to the eXist-db web application entitled "Bibliografia da Serra da Capivara".

References

ANDERSON, Tim (2001). *The XML Revolution. An interview with Charles Goldfarb. The XML Handbook.* [Consult. 4.04. 2014.]. Disponível em <URL:http://www.xmlhandbook.com/press/anderson.htm >

BASEX. (2014). *BaseX | The XML Database*. [Consult. 11.05.2014.]. Disponível em <URL:http://basex.org/>

BERNERS-LEE, Tim (1998). *The World Wide Web: A very short personal history. People of the W3C.* [Consult. 07.05. 2014.]. Disponível em <URL:http://www.w3.org/People/

Berners-Lee/ShortHistory.html.
CARROLL, Lewis (1917). *Through the Looking Glass, and What Alice Found There*. 232p, 99. Chicago: Rand McNally. [Consult. 11.05. 2014.]. Disponível em <URL:https://ia600309.us.archive.org/17/items/throughlookinggl00carr7/throughlookinggl00carr7.pdf.>
CHAPPLE, Mike (2014). *The ACID Model. About Databases.* [Consult. 12.05.2014.]. Disponível em <URL: http://databases.about.com/od/specificproducts/a/acid.htm.
CHECKLAND, Peter, and Sue HOLWELL (1997). *Information, Systems and Information Systems: Making Sense of the Field*. 278p. Chichester: John Wiley.
CHIPPINDALE, Christopher (1991). Editorial. *Antiquity*, 65 (248): 439–446.
CHIPPINDALE, Christopher (2000). Capta and Data: On the True Nature of Archaeological Information. *American Antiquity*, 65 (4): 605–612.
CODD, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13 (6):377–387. [Consult. 11.05.2014.]. Disponível em <URL:http://www.seas.upenn.edu/~zives/03f/cis550/codd.pdf.
CROCKFORD, Douglas (2008). *JSON. JavaScript: The Good Parts.* 172p, 136–145. Sebastopol: O'Reilly Media.
D'ABREO, Lara (2007). *DocBook: Write Once, Read Anywhere Documentation. Java Zone*. [Consult. 11.05.2014.]. Disponível em <URL: http://www.devx.com/Java/Article/35301.
DEAN, J. Claire (2006). Preserving a Worldwide Heritage: A Discussion about Rock Art Conservation. Conservation Perspectives, *The GCI Newsletter*, 21 (3):10–15. [Consult. 11.05.2014.]. Disponível em <URL: http://www.getty.edu/conservation/publications_resources/newsletters/pdf/v21n3.pdf.
DUBINKO, Micah (2003). XForms Essentials. 240p. Sebastopol: O'Reilly Media.
EVANS, Eric (2009). NOSQL 2009. Eric Evans' Weblog. [Consult. 12.05.2014.]. Disponível em <URL: http://blog.sym-link.com/2009/05/12/nosql_2009.html.
EXIST-DB (2014). eXist-db Open Source Native XML Database. [Consult. 11.05.2014.]. . http://exist-db.org/exist/apps/homepage/index.html.
FOSSATI, Angelo (1997). Rupestrian Archaeology. Tracce - Online rock art bulletin, 6. 8 January. [Consult. 11.05.2014.]. Disponível em <URL: http://www.rupestre.net/tracce/?p=1161.
GENNICK, Jonathan (2010). SQL Pocket Guide. 3rd ed. 206p. Sebastopol: O'Reilly Media.
GOLDFARB, Charles F. (1996). The Roots of SGML - A Personal Recollection. Charles F. Goldfarb's SGML Source Home Page. [Consult. 04.04.2014.]. http://www.sgmlsource.com/history/roots.htm.
HAROLD, Elliotte Rusty, and W. Scott MEANS. 2004. *XML in a Nutshell.* 714p. Sebastopol: O'Reilly Media.

HAUGEN, Knut (2010). *A Brief History of NoSQL. All About the Code*. 16 March. [Consult. 11.05.2014.]. Disponível em <URL: http://blog.knuthaugen.no/2010/03/a-brief-history-of-nosql.html.

JAFFE, Ludwig (2007). *Databases Without Databases for the World of Rock Art*. Conference of the British Rock Art Group, entitled, World of Rock Art, May 5–6. University of Cambridge McDonald Institute for Archaeological Research, Cambridge, UK. (presentation).

JAFFE, Ludwig (2010). *Captabases without Captabases for Rupestrian Imagery*. Meeting of the British Rock Art Group, May 8. University of Cambridge McDonald Institute for Archaeological Research, Cambridge, UK. (presentation).

Oxford Dictionaries. (2014). Rupestrian. Oxford Dictionaries. Rupestrian Oxford: Oxford University Press. [Consult. 11.05.2014.]. Disponível em <URL: http://www.oxforddictionaries.com/us/definition/english/rupestrian (11 May 2014).

ROBBINS, Jennifer Niederst (2010). *HTML & XHTML Pocket Reference*. 192p. Sebastopol: O'Reilly Media.

ROE, Charles 2012. *NoSQL or NoREL? A Short Account of Taxonomic Development. Learn Data Management at DATAVERSITY*. [Consult. 51.02.2014.]. Disponível em <URL: http://www.dataversity.net/nosql-or-norel-a-short-account-of-taxonomic-development/.

SIMPSON, John E. (2002). *XPath and XPointer: Locating Content in XML Documents*. 210p. Sebastopol: O'Reilly Media.

STROZZI, Carlo (2010). NoSQL. A Relational Database Management System. NoSQL Relational Database Management System. [Consult. 04.04.2014.]. Disponível em <URL: http://www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/NoSQL/.

STUBBLEBINE, Tony (2007). *Regular Expression Pocket Reference*. 128p. Sebastopol: O'Reilly Media.

TIDWELL, Doug (2008). *XSLT, Second Edition*. 990p. Sebastopol: O'Reilly Media.

UNIVERSITY OF CHICAGO PRESS STAFF 2010. *The Chicago Manual of Style*, 16th edition. Chicago: Chicago University Press.

WALMSLEY, Priscilla (2007). *XQuery*. 512p. Sebastopol: O'Reilly Media.

WALSH, Norman (2010). *DocBook 5: The Definitive Guide*. 552p. Sebastopol: O'Reilly Media.