Daniel Gomes - Full CV

Expert on Web Archiving and Web-based information systems, Product Manager, Technical Team Leader

Last update in November 2023

Contacts

- danielcoelhogomes@gmail.com, +351 91 226 92 87
- LinkedIn
- Google Scholar
- Professional page

Short biography

Daniel Gomes started Arquivo.pt (the Portuguese web-archive) in 2007 as follow-up of his PhD thesis and currently leads this public service that provides worldwide unique features (e.g. image and textual search over information archived from the web since the 1990s).

The Arquivo.pt software architecture is composed of 8 systems that host 35 components which manage over 1 PetaByte of data (20 billion web files) distributed over 80 servers at its own data center. The current software system is the result of more than 25 major releases since 2015.

Daniel Gomes dedicated most of his career to developing the Arquivo.pt public service. For this purpose, he developed a wide scope of expertises along time because a web archive is a peculiar information system which is not supported by major suppliers nor web archiving is yet an established area of knowledge.

He obtained his PhD in Computer Science with a thesis focused on the design of large-scale systems for the processing of web data to automatically derive evolution trends (a.k.a. Big Data).

He has been working in the research, development and operation of web-based information systems since 2001, covering all stages of the lifecycle of a product from theoretical scientific hypotheses to services successfully running for years. He has managed software teams with frequent staff rotation and open-source projects since 2007.

He delivered over 200 invited talks and training sessions and published over 57 scientific and technical documents. He was the Chief Editor of the book “The Past Web: Exploring Web Archives” and independent expert to evaluate research activities funded by the European Commission on Information and Communication Technologies.

As strength, he highlights the ability to combine scientific methodologies with pragmatism to make decisions. As weakness, lack of tolerance for unproductive bureaucratic processes.

Career of Daniel Gomes summarized in a word cloud.

Skills

- Long-term experience in the development and management of all stages of web archiving and web-based information systems, from research ideas until they

become high-availability services (e.g. Arquivo.pt was idealized as a research hypothesis and it is a robust public service running since 2012).
- Remote management of distributed teams with high staff rotation (e.g. managed service for 15 years with one team member rotation every 8 months, on average).
- Pragmatic solutions to complex problems by applying scientific methodologies.
- Advocacy, communication and training with proficiency in varied writing styles (e.g. web, scientific, formal, marketing, creative).
- Close connections to academic and research organizations in the field of Computer Science that facilitate the establishment of collaboration projects and hiring of staff.

Expertise

- Web-based information systems
- Web crawling, archiving, search and digital preservation
- Information retrieval, Natural Language Processing and Machine Learning
- Big Data, Distributed and Cloud systems
- Site Reliability Engineering (SRE), Quality assurance and Cybersecurity
- User Experience, Design Thinking, Web Accessibility and Usability
- Management of software teams, projects and products
- Recruitment, training and coaching
- Advocacy, communication and dissemination

Areas of interest

- Ecology
- Cognitive Psychology
- Behavioral Economics

Favorite Technologies

- Open-source platforms based on Java/Linux
- Big Data and Search: Hadoop, Apache Spark, Apache Tika, Lucene/Solr, Nutch
- Databases: PostgreSQL, Oracle
- Web publishing: HTML, CSS, Javascript, PHP, Wordpress, Tomcat, HTTPd
- Web archiving: Internet Archive (Heritrix, Brozzler, Wayback Machine), webrecorder.net tool suite (archivewebpage, pywb, browsertrix)
- SRE and DevOps: Python, Ansible, Puppet, SonarQube/SonarCloud, SauceLabs, Jenkins, SelleniumHQ, ICINGA, Awstats, Jmeter
- Management and collaboration: Google Workspace, GitHub, Slack, Atlassian tools (Jira, BitBucket, Confluence, Trello)

Personal skills

- Leadership, responsibility and resilience
- Concise communication adapted to message receivers
- Pragmatic, data-driven and objective-oriented decision-making
- Time management and organization
- Critical analysis and strategic planning

- Initiative, continuous learning and strive for excellence
- Mission-oriented team spirit

Languages

- Portuguese: fluently written and spoken
- English: fluently written and spoken
- Spanish: good understanding of written and reasonably spoken
- French: good understanding of written, reasonably spoken

Education

- PhD in Computer Science by the University of Lisbon (2007), Web Modelling for Web Warehouse Design, (ppt). Advisor: Mário J. Silva, Jury: Ricardo Baeza-Yates et al.
- Graduated in the Faculty of Sciences of University of Lisbon, specialized in Information Systems, second-best of his course (2001).
- Training: Usability and user experience (Norman Nielsen Group), Oracle databases, finances, procurement, negotiation, pedagogical techniques, project management, interpersonal relations, production of scientific literature, attended over 100 scientific and technical events.

Honors and awards

- Best Digital Service of 2022 for Arquivo.pt.
- Best paper award for Arquivo e medição da web portuguesa at Ibero-Americana IADIS WWW/Internet 2008.
- Best paper by young researcher for Design and Selection Criteria for a National Web Archive at Theory and Practice on Digital Libraries 2006.

Experience

- **Head of the Arquivo.pt** public service at the Foundation for Science and Technology which is a governmental institution (2013-current). The core of his team has been composed on average by 5 members with varied profiles (software engineers, communication professionals, librarians, researchers) but the majority held MSc and PhD in Computer Science. The main activities and achievements were:
    - Promoted over 180 dissemination activities.
    - Launched the largest search service over web-archived images.
    - Developed and operated a catalog of 13 world-wide services complementary to web archives (e.g. SavePageNow, arquivo404 and Memorial).
    - Conducted Site Reliability Engineering (SRE) and redesigned the system to a shared-nothing architecture to transition from project to service (availability over 99,99% since 2016).
    - Published derived open data sets for research.
    - Launched and managed the Arquivo.pt Awards since 2018 (167 applications, 23 awarded).
    - Produced pedagogical materials and designed the training program on web preservation and research delivered in-person and remotely (1 289 trainees with average satisfaction of 89% in 2022).

  - Designed the Arquivo.pt Application Programming Interfaces.
  - Procurement for obtaining hardware and software resources.
  - Representative of Arquivo.pt at the International Internet Preservation Consortium.
  - Arquivo.pt preserved the online documentation about European and national scientific projects.
  - Arquivo.pt was considered the best Digital Service of 2022.
  - Arquivo.pt entered the honor roll for cibersecurity in Portugal.
- **Invited jury** of MSc and PhD theses at several universities and member of Programme Committee for scientific publications and events about web archiving, digital libraries and information retrieval (2005-current).
- **Chief Editor** of the book "The Past Web: Exploring Web Archives" published in 2021 (27 000 accesses).
- **Technical team leader** of the web development and operation team at FCCN Foundation for National Scientific Computing (FCCN) which was a private non-profit organization (2011-2015). His team was composed on average by 10 members (software engineers and designers). The main tasks performed were project and team management, usability testing, quality assurance and system administration. His team contributed to the design, development and maintenance of over 20 projects and services such as:
  - Open access initiative for scientific repositories (RCAAP)
  - Online Knowledge Library (B-on)
  - Video repositories for Research and Education: Educast and Zappiens
  - Program to bridge universities with schools: Mundo na Escola (World at School)
  - Online service to measure network speed (ANACOM Netmede.net)
  - Informational websites (e.g. Harvard Medical School Portugal, Study & Research in Portugal, Speedmeter, FCCN.pt, Jornadas FCCN, Natureza na Web, ISOC Portugal)
  - Portuguese web archive project
- **Researcher/Project manager at FCCN** (2007-2011). Main tasks: research, system administration (Linux operating systems, Virtual Machines, Storage Area Networks, databases) and management of the "Portuguese Web Archive" project which originated the first open-source full-text search prototype over a web archive and the recommendations for publishing preservable web content
- **Independent expert** on Search Engine Optimization, web-based information systems, accessibility and usability (2007-2014).
- **Independent expert** to evaluate research activities funded by the European Commission on Information and Communication Technologies (2009-2012)
- **Author of Visibilidade.net**, a website in Portuguese about Search Engine Optimization, Usability and Accessibility for people with disabilities (2006-2012)
- **Researcher** at the Large-Scale Information Systems Laboratory/PhD student (2004-2007). Mainly studied web crawling, data processing and information retrieval (a.k.a. Data Science). He developed the first prototype of a Portuguese web archive and a distributed system to process large amounts of web data (Chapter 5 of Web Modelling for Web Warehouse Design).

- **Project manager** in a startup company focused on Web publishing: search engine optimization, usability, accessibility for people with disabilities and web marketing (2007-2008).
- **System/database administrator** at LaSIGE (2002-2004): Linux, LAN, DNS, NFS, Oracle and PostgreSQL databases used to process tables with millions of rows to derive scientific results from web data.
- **Developer/system administrator/researcher** at ICAT Institute of Applied Science and Technology (2000-2002). Participated in the development of the first Portuguese news search engine that enabled theme-based queries across distinct online publications (Newssearch) and developed the middleware for the first mobile leisure guide supported by Palm Pilot PDAs.

Project engagement

- ROSSIO Infrastructure is an access point to free and open content on social sciences, arts and humanities (2014- present), responsible for scientific reviewing, training and management.
- RESAW.eu (2012-present): Research Infrastructure for the Study of Archived Web Materials, responsible for scientific reviewing, training and management.
- CLEOPATRA (2019-2023): Cross-lingual Event-centric Open Analytics Research Academy, responsible for scientific reviewing, training and management.
- Linguateca (1998-2020): distributed resource center for the Portuguese language, responsible for research, web crawling and creation of research corpora.
- QualWeb (2008-2011): automatic evaluation of the web accessibility for people with disabilities, research, management and scientific reviewing.
- Tomba (2005-2006): first prototype of a Portuguese web archive, responsible for the web warehousing system and the technical coordination of the project.
- XMLBase (2002-2004): researched methods for systems that manage semi-structured data, responsible for the design and development of the warehousing framework.
- Tumba! (2000-2006): search engine optimized for the Portuguese web, responsible for the design, development and maintenance of the crawling and web warehousing systems.
- Digital Deposit (2000-2001): studied the design of a system to archive selected online publications, responsible for the design and development of the web data collection component.

Publications peer-reviewed

- André Mourão, Daniel Gomes, Searching images in a web archive, 10th IEEE International Conference on Data Science and Advanced Analytics 2023.
- Daniel Gomes, Web archives as research infrastructure for digital societies: the case study of Arquivo.pt, 2022.
- Sawood Alam, Michele C. Weigle, Michael L. Nelson, Fernando Melo, Daniel Bicho, Daniel Gomes, MementoMap Framework for Flexible and Adaptive Web Archive Profiling, Joint Conference on Digital Libraries 2019 (ppt, abstract).

- Helen Hockx-Yu, Ditte Laursen, Daniel Gomes, The Curious Case of Archiving .eu, The Historical Web and Digital Humanities: The Case of National Web Domains, 2019.
- Daniel Bicho, Daniel Gomes, Preserving Websites Of Research & Development Projects, International Conference on Digital Preservation, 2016 (ppt).
- Miguel Costa, Daniel Gomes, Mário J. Silva, The Evolution of Web Archiving, International Journal of Digital Library Systems, 2016.
- Daniel Gomes and Miguel Costa, The Importance of Web Archives for Humanities, International Journal of Humanities and Arts Computing, 2014.

- Daniel Bicho, Daniel Gomes, Preserving Websites Of Research & Development Projects, International Conference on Digital Preservation, 2016 (ppt).
- Daniel Gomes and Miguel Costa, The Importance of Web Archives for Humanities, International Journal of Humanities and Arts Computing, 2014.

- Daniel Gomes, David Cruz, João Miranda, Miguel Costa, Simão Fontes, Acquiring and providing access to historical web collections (demo), 10th International Conference on Preservation of Digital Objects, 2013 (ppt).
- David Cruz, Daniel Gomes, Adapting search user interfaces to web archives, 10th International Conference on Preservation of Digital Objects, 2013 (ppt, poster).
- Miguel Costa, João Miranda, David Cruz, Daniel Gomes, Query suggestion for web archive search, 10th International Conference on Preservation of Digital Objects 2013 (ppt, poster).
- Miguel Costa, Daniel Gomes, Francisco M Couto, Mário J. Silva, A Survey of Web Archive Search Architectures, Temporal Web Analytics Workshop 2013 (ppt).
- Daniel Gomes, Miguel Costa, David Cruz, João Miranda, Simão Fontes, Creating a Billion-Scale Searchable Web Archive, Temporal Web Analytics Workshop 2013 (ppt).
- Daniel Gomes, David Cruz, João Miranda, Miguel Costa, Simão Fontes, Search the Past with the Portuguese Web Archive, 22nd International World Wide Web Conference, 2013 (poster, ppt).
- Daniel Gomes, João Miranda, Miguel Costa, A survey on web archiving initiatives, International Conference on Theory and Practice of Digital Libraries 2011 (ppt, poster, video).
- Rui Lopes, Daniel Gomes, Luís Carriço, Web Not For All: A Large Scale Study of Web Accessibility, W4A: 7th ACM International Cross-Disciplinary Conference on Web Accessibility, 2010 (ppt).
- João Miranda, Daniel Gomes, Trends in Web characteristics, 7th Latin American Web Congress, 2009.
- João Miranda, Daniel Gomes, An Updated Portrait of the Portuguese Web, 14th Portuguese Conference on Artificial Intelligence (EPIA 2009).
- João Miranda, Daniel Gomes, How are Web characteristics evolving (poster), ACM Hypertext Conference, 2009.
- Daniel Gomes, João Miranda, Arquivo e medição da web portuguesa (best paper award), Ibero-Americana IADIS WWW/Internet 2008.
- Daniel Gomes, Mário J. Silva, The Viúva Negra crawler: an experience report, Software: Practice and Experience, 2008.
- Daniel Gomes, André Nogueira, João Miranda, Miguel Costa, Introducing the Portuguese web archive initiative, 8th International Web Archiving Workshop, 2008 (ppt).

- Daniel Gomes, Sérgio Freitas, Mário J. Silva, Design and Selection Criteria for a National Web Archive (best paper by young researcher award), 10th European Conference on Research and Advanced Technology for Digital Libraries 2006 (ppt).
- Daniel Gomes, Mário J. Silva, Modelling Information Persistence on the Web (best paper candidate), The Sixth International Conference on Web Engineering, 2006 (ppt).
- Nuno Cardoso, Bruno Martins, Daniel Gomes, Mário J. Silva. WPT 03: a primeira colecção pública proveniente de uma recolha da web portuguesa, Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa, 2007.
- Daniel Gomes, Mário J. Silva, Characterizing a National Community Web, ACM Transactions on Internet Technology, 2005.
- Daniel Gomes, André Santos, Mário J. Silva, Managing duplicates in a web archive, 21th Annual ACM Symposium on Applied Computing, 2006 (ppt).
- Norman Noronha, João P. Campos, Daniel Gomes, Mário J. Silva, José Borbinha, A Deposit for Digital Collections, Fifth European Conference on Research and Advanced Technology for Digital Libraries, 2001.
- Daniel Gomes, André Nogueira, João Miranda, Miguel Costa, Introducing the Portuguese web archive initiative, 8th International Web Archiving Workshop, 2008 (ppt).
- Daniel Gomes e Mário J. Silva, A Characterization of the Portuguese Web, 2003 (ppt).
- Daniel Gomes, Mário J. Silva, Tarântula - Sistema de Recolha de Documentos da Web, 2001.
- Daniel Gomes, João Campos, Mário J. Silva, Versus: A Web Repository, 2002.

Selected gray literature

Reports

- Daniel Bicho, Fernando Melo e Daniel Gomes, An Evaluation of Replay Quality for Web-Archived Pages, 2017.
- Fernando Melo, Hugo Viana, Daniel Gomes, Miguel Costa, Architecture of The Portuguese Web Archive Search System version 2, 2016.
- Fernando Melo, Daniel Gomes, Relatório de Testes de Usabilidade à ReplayBar do site Arquivo.pt, 2016.
- Fernando Melo, Daniel Bicho e Daniel Gomes, A Comparison Between The Performance of Wayback Machines, 2016 (ppt).
- Hugo Viana, Daniel Gomes e Miguel Costa, Architecture of The Portuguese Web Archive Search System, 2016 (ppt).
- Daniel Bicho e Daniel Gomes, Estudo de gestão de capacidade do Arquivo.pt, 2015.
- Daniel Bicho, João Miranda, Daniel Gomes, A first attempt to archive the .EU domain, 2015.
- David Cruz and Daniel Gomes, Tweaking search user interfaces to web archives, 2013.
- Daniel Gomes, David Cruz, João Miranda, Miguel Costa e Simão Fontes, Creating a searchable web archive, 2012.

- Miguel Costa, João Miranda, David Cruz e Daniel Gomes, Query Suggestion for Web Archive Search, 2012.
- Daniel Gomes, Tarântula - Sistema de Recolha de Documentos na WWW, 2001.
- Carlos Correia, Daniel Gomes, Norman Noronha, Guia Móvel do Lazer: Planeamento, Análise e Desenho, 2000.
- Carlos Correia, Daniel Gomes, Norman Noronha, Interface de Utilização do NewsSearch, 2000.

Talks and interviews

- Sociedades digitais sem memória? (Memoryless digital societies?), 2023
- Arquivo.pt behind the curtains, 2023
- How to research governmental web data? (slides), 2023
- Tutorial: how to explore Arquivo.pt using Python, 2022
- 15 anos de Arquivo.pt contados por Daniel Gomes (15 years of Arquivo.pt told by Daniel Gomes), 2022
- Catálogo de serviços do Arquivo.pt (Catalog of Arquivo.pt services), 2022
- Arquivo.pt as open data provider, 2022
- Challenges and Recommendations to start a web archive, 2021
- Arquivo.pt: o que é? Para que serve? (Arquivo.pt: What? What for?, 2021
- Arquivo.pt (podcast 1/0), 2020
- Buscando el pasado con Arquivo.pt, (in Spanish) (Searching the past with Arquivo.pt), 2020
- Access policies, challenges and approaches (slides), 2019
- Tutorial: Research the Past Web using Web archives, 2019
- Tutorial sobre investigação utilizando arquivos da web (Tutorial on researching web archives), 2018
- Arquivo.pt: a case study interview, 2019
- Arquivo.pt e preservação do conteúdo brasileiro (Preservation of Brazilian web content), 2019
- Arquivo.pt Memorial and other goodies, 2019
- Novas formas de pesquisar o passado: 10 anos do Arquivo.pt (New ways of searching the past: 10 years of Arquivo.pt), 2017
- Improving the robustness of the Arquivo.pt web archive, 2016
- Preservation of web information about the Charlie Hebdo attack, 2016
- Web Archive Information Retrieval, 2015
- Wiki não é uma tecnologia, é uma atitude (Wiki is not a technology, is an attitude), 2014
- Análise de requisitos e controlo de qualidade no desenvolvimento de sites – Uma cábula rápida (Cheatsheet for requirement analysis and quality assurance on web development), 2013

Civic activities and hobbies

- Surfer, environmental activist and defender of animals rights.
- He founded the TaraRecuperavel.org civic movement which acts to improve environmental awareness in Portugal and collaborated with ecological non

governmental organizations (Quercus, Brigada do Mar, APLM - Portuguese Association for Marine Litter).

    - Designed, managed and executed Guerrilla Marketing campaigns on social networks and developed over 148 ecological awareness activities and documents.
    - Contributed and lobbied for the approval of environmental laws:
        - Law 69/2018: incentive system for the return and deposit of plastic, glass, ferrous metal and aluminum beverage containers (bottle bills).
        - Law 76/2019: determines the non-use and non-availability of single-use plastic tableware in the restaurant and/or beverage sector and in the retail trade.
        - Law 88/2019: reduction of the impact of cigarette butts, cigars or other cigarettes on the environment.
- PADI Advanced Open Water Diver.