João D. **Ferreira**

Computer Scientist · Bioinformatician

R. das Flores, nº 7 – 1º dto, 2810-218 Laranjeiro, Portugal

(+351) 96 323 50 69 | jdferreira@fc.ul.pt | http://webpages.fc.ul.pt/~jdferreira

ORCID 0000-0002-6900-3168 | jdferreira | jdferreira

Summary

Professional Experience

2015–Present	Professor Auxiliar Convidado (50%) *at Faculdade de Ciências da Universidade de Lisboa*
2016–Present	Collaborator *at Laboratório de Sistemas de Informação de Grande Escala*
2012–2014	Hired Researcher *by the EPIWORK project*

Scientific Performance

Education	Ph.D., Computer Science, Universidade de Lisboa
Publications	6 peer-reviewed journal publications 6 conference and workshop publications total citations: 92; *h*-index: 6 (from Google Scholar)
Software	3 open-source software programmes 1 online web application
Knowledge resources	2 knowledge-representation artefacts
Projects	participation in 2 national projects and 1 European project
Academic Supervision	1 M.Sc. in Informatics Engineering
Conference organization	ICBO 2015 Workshop & tutorial Chair ICBO 2015 Session chair
Reviewer	3 international journals 2 international conferences 2 national conferences
Honors and Awards	2 scholarships (PhD and MSc)

Pedagogical Performance

Taught courses	Aplicações e Serviços na Web Aplicações na Web Bioinformática Bioinformatics & Computational Modelling Interacção com Computadores Introdução às Tecnologias da Web Ontologias Aplicadas às Ciências Processamento de Dados Programação
Pedagogical Surveys	Average of 3.78 out of 4

A Scientific performance

A.1 Scientific production

Selected list of publications

[1] Andre Lamurias, **João D. Ferreira**, and Francisco M. Couto. "Improving chemical entity recognition through h-index based semantic similarity." In: *Journal of Cheminformatics* 7.Suppl 1 Text mining for chemistry and the CHEMDNER track (Jan. 2015), S13. ISSN: 1758-2946. DOI: 10.1186/1758-2946-7-S1-S13

- **Impact Factor**: 4.550
- **Non-self citations**: 4

[3] Catia Pesquita, **João D. Ferreira**, Francisco M. Couto, and Mário J. Silva. "The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources." In: *Journal of Biomedical Semantics* 5.1 (2014), p. 4. ISSN: 2041-1480. DOI: 10.1186/2041-1480-5-4

- **Impact Factor**: 2.262
- **Non-self citations**: 2

[4] **João D. Ferreira**, Janna Hastings, and Francisco M. Couto. "Exploiting disjointness axioms to improve semantic similarity measures." In: *Bioinformatics* 29.21 (2013), pp. 2781–2787. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt491

- **Impact Factor**: 4.981
- **Non-self citations**: 9

[5] **João D. Ferreira**, Daniela Paolotti, Francisco M. Couto, and Mário J. Silva. "On the usefulness of ontologies in epidemiology research and practice." In: *Journal of epidemiology and community health* (Nov. 2012), pp. –3. ISSN: 1470-2738. DOI: 10.1136/jech-2012-201142

- **Impact Factor**: 3.501
- **Non-self citations**: 5

[6] **João D. Ferreira** and Francisco M. Couto. "Semantic Similarity for Automatic Classification of Chemical Compounds". In: *PLoS Computational Biology* 6.9 (Sept. 2010). Ed. by John B. O. Mitchell, e1000937. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000937

- **Impact Factor**: 4.620
- **Non-self citations**: 33

Histograms of publications

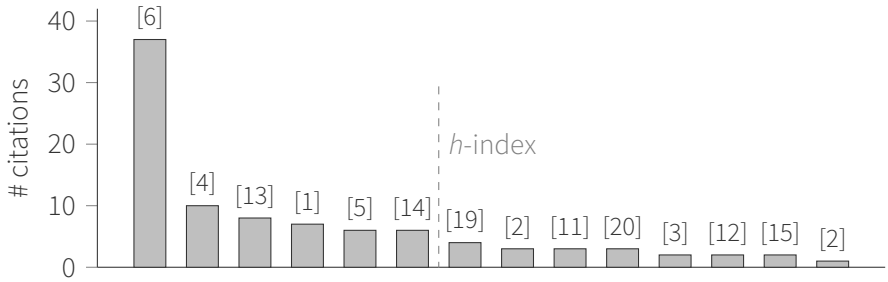

Figure 1: Histogram of citations for each of my cited papers, with papers sorted from most cited to least cited. Top labels specify the publication according to Section E. The vertical dashed line shows the associated *h*-index.

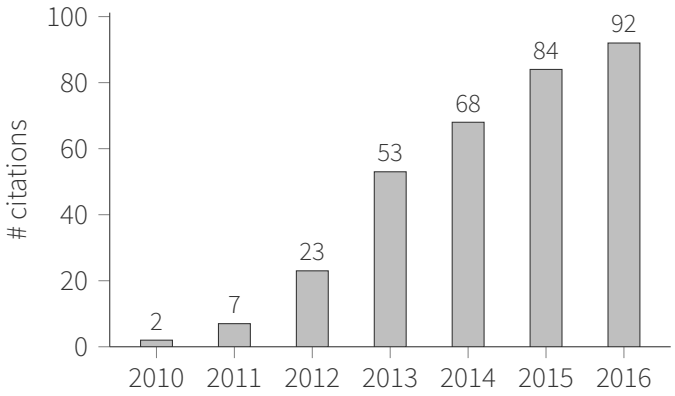

Figure 2: Histogram showing the cumulative number of citations gathered throughout the years.

Selected list of citations

Stefan Schulz et al. "From concept representations to ontologies: a paradigm shift in health informatics?" In: *Healthcare informatics research* 19.4 (2013), pp. 235–242

- Mentions my paper [5] as **one of the twenty most influential papers** in the areas of knowledge-representation, biomedical ontologies and electronic health issues, as listed by the members of the LinkedIn Working Group on Medical Concept Representation.

Janna Hastings et al. "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013". In: *Nucleic acids research* 41.D1 (2013), pp. D456–D463

- This paper is a major milestone in Cheminformatics, describing ChEBI (an ontology of chemical compounds). It has itself 163 citations. Cites my paper [6] as an example of an **application that makes use of the semantic information** encoded in ChEBI for automatic classification.

Janna Hastings et al. "Structure-based classification and ontology in chemistry." In: *J. Cheminformatics* 4 (2012), p. 8

- This paper expresses **the need and usefulness of systems that exploit the machine-readable information** provided by reference ontologies, and cites my paper [6] as an example.

Joerg Kurt Wegner et al. "Cheminformatics". In: *Communications of the ACM* 55.11 (2012), pp. 65–75

- This paper advocates the use of **open-source solutions to deal with the vast amount of existing chemical information**, mentioning semantic similarity in ChEBI and citing my paper [6].

Robert Hoehndorf et al. "Thematic series on biomedical ontologies in JBMS: challenges and new directions." In: *J. Biomedical Semantics* 5 (2014), p. 15

- This paper aims "to disseminate the latest developments in research on biomedical ontologies and provide a venue for publishing newly developed ontologies, updates to existing ontologies as well as methodological advances, and selected contributions from conferences and workshops". It cites my paper [6] as an example of a semantic similarity application and my paper [4] as **the start of using more than taxonomies in semantic similarity** in the biomedical domain.

Open Software

OWLtoSQL

2015

URL: `http://github.com/jdferreira/owltosql`

- This software is one of the results of my PhD work. It is used to **convert an ontology written in OWL** (Web Ontology Language) **into a relational database**, which offers several advantages, the most promi-

nent of which is that it enables *random* access to the ontology constituents (classes, properties, axioms, *etc.*). It has been used by Bruno Inácio in his M.Sc work (see Section A.3).

Multi-domain Ontology-based Semantic Similarity (MOSSy) *2015*

URL: `http://github.com/jdferreira/mossy`

- This software also resulted from my PhD work. It is responsible for calculating **semantic similarity between resources annotated with concepts from multiple ontologies**. The program is configurable to use any OWL ontology and can handle any type of annotated resource, both in single- and multi-domain contexts. It is extensible, since it allows the implementation of semantic similarity measures as Python classes.

Annota *2010*

URL: `https://code.google.com/archive/p/annota/`

- This software was used internally in the GREASE-II project to assist the **manual disambiguation** of ambiguous tags automatically assigned to terms in free-text news reports.

Web applications

MOSSy *2015*

URL: `http://lasige.di.fc.ul.pt/webtools/mossy`

- This **web application is a gateway to apply MOSSy** (the program, see above) to a collection of annotated resources. The application first requests from the user a set of models which can be given in free text or already annotated resources. If free-text models are provided, a text-mining service is used to extract annotations. Given the set of models, the application compares them according and outputs a cluster of the resources as well as the similarity matrix between them.
  The software was used by a research team in University College London to showcase similarity between physiological models of tissue, but is available for use by the whole scientific community.

Knowledge representation artefacts

Network of Epidemiology-Related Ontologies *2013*

See my paper [5]

- This work is a **collection of ontologies related to the epidemiology domain**. It contains 13 ontologies, from domains such as biochemistry, diseases, environment, transmission, vaccines, and geography, which provide the necessary concepts to annotate a digital resource of the epidemiological domain. The network was later included in the Epimarketplace, a repository of epidemic resources, in order to facilitate the annotation process.
- **Developed**: during my work in the EPIWORK project.

Epidemiology Ontology *2014*

URL: `https://code.google.com/p/epidemiology-ontology/`

- This ontology represents both **disease transmission methods and epidemiology parameters**, containing over 300 concepts. This ontology fills the previously existing gap in these areas, which were under-represented in biomedical ontologies. It supports the semantic annotation of resources.
- **Developed**: during my work in the EPIWORK project.

A.2 Research projects

Past projects

Semantic Ontology Matching using External Resources (SOMER)

2012-2014

URL: `http://somer.di.fc.ul.pt/`

- **Position**: Member of research team
- **Funding**: FCT (84 000€)
- **Grant**: PTDC/EIA-EIA/119119/2010
- **Description of project**: The SOMER project aimed to develop ontology matching methods that exploit external knowledge resources through evidence and information content provided by unstructured text and annotation corpora. The results were applied to real world applications based on existing ontologies in the biomedical and geospatial areas.
- **My contributions**: I was involved in writing the proposal for this project. I was also responsible for exploring ways that allowed the use of semantic similarity measures in the process of ontology matching. I contributed to 1 peer-reviewed publication.

Developing the Framework for an Epidemic Forecast Infrastructure (EPIWORK)

2012-2014

- **Position**: Hired researcher
- **Funding**: FP7 (5 000 000€)
- **Grant**: 231807
- **Description of project**: The EPIWORK project proposed a framework of tools and knowledge for the design of epidemic forecast infrastructures, including: development of the mathematical and computational methods needed to achieve prediction of disease spreading in complex social systems; development of large scale, data driven computational models aimed at epidemic scenario forecast; design and implementation of data-collection mechanisms, such as the collection of real-time disease incidence, through innovative web and ICT applications; and the implementation of a computational platform for epidemic research and data sharing.
- **My contributions**: The LASIGE team was responsible for the creation go the Epidemic Marketplace, an online platform for epidemic research and data sharing was the data-hub for the multiple research communities and countries involved in the project. I was hired by the LASIGE team as a consultant in semantic web, specifically to design the semantic metadata model that was used to annotate the resources of the marketplace, as well as to design the Network of Epidemiology-Related Ontologies, which contributed to the repository as a collection of concepts used to annotate the resources. I contributed to 5 peer-reviewed publications and 1 open-source ontology.

Geographic Reasoning for Search Engines II (GREASE-II)

2010

- **Position**: M.Sc fellowship
- **Funding**: FCT (117 300 €)
- **Grant**: PTDC/EIA/73614/2006
- **Description of project**: This project aimed at researching information access methods to large collections of documents and objects having geographically rich text and meta-data. One of the main ideas of the project was that geospatial information can be mapped into ontology concepts.
- **My contributions**: One of the tasks of this project was to perform text-mining directly onto news reports in order to find geographical references in text; to assist in this task, a disambiguation module had to be created. I was assigned the creation of software to help disambiguate annotations manually, which would then later be used in a machine-learning step. I was also tasked with the alignment of Geo-Net-PT (a geographical ontology of the Portuguese territory) to Yahoo! GeoPlanet™ (an infrastructure for geo-referencing data on the Internet). I contributed to 1 technical report and to 1 open-source software

Currently being considered for funding

Plataforma de monitorização alimentar para controlo inteligente de nutrição personalizada *2016*

- **Description of the project**: This project has been recently submitted to a Portugal 2020 call and as such its technical details cannot be fully disclosed. One of the ideas that will be implemented, however, is the use of semantic data concerning food and nutrition to help individuals improve their food intake.
- **My contributions**: Up until this point, my contribution to the project was writing the proposal in collaboration with the rest of the research and business teams. I was responsible for a significant part of the state-of-the-art concerning semantic data on the web, as well as for the description of the tasks related to semantic-web technologies.

International Collaborations

University College London (UCL) *2013-ongoing*

- I collaborated with Bernard D. Bono, a Principal Research Fellow in Health Informatics in UCL, in several fronts. The most recent collaboration is related to the **digital representation of functional tissue units** (small portions of an organ that perform a certain biological function) and the ways to compare them and publish them in a knowledge-base. This collaboration resulted in an oral presentation at the Virtual Physiological Human Conference 2014 [7].

European Bioinformatics Institute (EBI) *2013-2016*

- I collaborated with J. Hastings, former group coordinator of the Cheminformatics and Metabolism Team at EBI, in devising a semantic similarity measure that can effectively **capture the information provided by the disjointness axioms of an ontology**. This collaboration resulted in 1 peer-reviewed publication [4].
- I am also currently collaborating with the team behind Metabolights, a repository used to store details of experiments related to research in metabolism and information derived from the research. My role in this collaboration is to help **devise ways to measure annotation quality** in order find resources under-annotated and to help authors submit properly annotated data. This collaboration is still ongoing and is expected to result in at least 1 peer-reviewed publication.

Institute for Scientific Interchange (ISI) Foundation *2012*

- I collaborated with Daniela Paolotti, the Project manager of the Italian web platform for Influenza-like Illness Surveillance. This collaboration resulted from my position at EPIWORK, and resulted in the publication of 1 peer-reviewed publication detailing **how to introduce semantic web technologies and ideas in the epidemiology domain** [5].

A.3 Academic supervision

Current supervision

M.Sc in Informatics Engineering *2015-2016*

Universidade de Lisboa

- **Student**: Bruno Inácio
- **Position**: Co-supervision with Francisco M. Couto, Universidade de Lisboa
- **Thesis**: "Quanto valem os metadados?"
- **Topic**: The work focussed (i) on the study of the quality of metadata as a way to sustain semantic integration and to facilitate data-sharing, based on the specificity of the ontology concepts used to annotate the resources and on the thoroughness of these annotations; and (ii) on the development of a platform that allows the assessment of metadata quality of the resources in a scientific data repository.

Internship of a B.Sc in Biochemistry *2011-2012*

Universidade de Lisboa

- **Student**: Hugo Ferreira
- **Position**: Co-supervision with Francisco M. Couto, Universidade de Lisboa
- **Topic**: Creation of an ontology of epidemiology.
- **Topic**: The work consisted in using the textual descriptions in a Dictionary of Epidemiology to find relationships between the concepts and thus create an ontology of epidemiology.

A.4 Grants, Honors and Awards

Grants

2010–2015 PhD Grant by Fundação para a Ciência e Tecnologia
2009–2010 Pre-PhD grant by LASIGE

Awards

2014 BPH Travel Award by the VPH Institute
2006 Best Student in 1st year by Universidade de Lisboa
2006 Merit Scholarship by Universidade de Lisboa

Conference organisation

6th International Conference on Biomedical Ontology (ICBO) *2015*

Workshop & tutorial chair, Proceedings chair

- **Description**: The sixth International Conference on Biomedical Ontology (ICBO) was held in Lisbon in 2015 (`http://icbo2015.fc.ul.pt`). This prestigious and well-attended conference gathered multidisciplinary researchers at the Faculdade de Ciências da Universidade de Lisboa to present and discuss the latest research breakthroughs in exploring ontologies in a biomedical and clinical context.
- **Position**: Part of the organization: I was Workshop & tutorial chair and Session chair. I was also part of the Proceedings team, having compiled the Proceedings, and organized the submission to CEUR-WS.

Reviewer

2016 Journal of Biomedical Semantics
2015 Journal of Biomedical Semantics
Journal of Epidemiology & Community Health
2013 Oxford Bioinformatics
2012 Oxford Bioinformatics

Program Committee and other reviewer activities

2015 Bioinformatics Open Days
Conference on Practical Applications of Computational Biology & Bioinformatics
Portuguese Conference on Artificial Intelligence
2014 Conference on Practical Applications of Computational Biology & Bioinformatics
International Conference on Data Integration & Life Sciences

B Pedagogical performance

B.1 Teaching

Courses

Table 1: This table shows my contribution to courses from the Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa.
Legend: *PL*: Prática de Laboratório (Practice lesson); *TP*: Teórico-Prática (Theoretical & Practical lesson)

Course	Position	2015–2016
ASW	PAC	**PL**
IC	PAC	**TP & PL**
ITW	PAC	**TP & PL**
PD	PAC	**TP**
Prog. I	PAC	**TP**

ASW (Aplicações e Serviços na Web) *2015–2016*

- **Degree**: Licenciatura em Tecnologias de Informação (2nd year)
- **My lectures**: PL
- **Topics**:
  - Web application characteristics and features;
  - The development process of web applications;
  - Introduction to the main server-side web technologies: resource addressing, protocols and general architecture;
  - The various data transfer formats (XML, JSON, *etc.*) and related technologies;
  - Introduction to Web Services and Semantic Web.
- **My contributions**: I produced a guide to explain HTML forms, along with JavaScript access to the content of the forms and the content of the GET arguments passed with the URL. The goal was to teach students how to deal with user-input, and how it travels through the network between web pages.

IC (Interacção com Computadores) *2015–2016*

- **Degree**: Licenciatura em Tecnologias de Informação (2nd year)
- **My lectures**: TP & PL
- **Topics**:
  - Introduction to Human-Computer Interaction (HCI);
  - The foundations of HCI: Human and technological aspects;
  - The design process: user centred design, interaction design basics, guidelines for interaction design, and evaluation techniques;
  - Models and theories: cognitive models, task analysis, dialogue notations
- **My contributions**: I contributed to the class materials produced, including one of the project descriptions, which consisted in the development of an application based on web technologies that assisted people managing their kitchen (kitchen utensils, silverware, food *etc.*). The goal was to teach students how to implement an application following the principles of human-computer interaction.

ITW (Introdução às Tecnologias da Web) *2015–2016*

- **Degree**: Licenciatura em Tecnologias de Informação (1st year)
- **My lectures**: TP & PL
- **Topics**:
  - Internet and web history;
  - Basic web concepts: the architecture, models, protocols, user agents, and transactions;
  - Text and Hypertext Mark-up;
  - Image Mark-up;
  - Introduction to HTML;

  - Introduction to Cascading Style Sheets (CSS);
  - JavaScript concepts: flow control, data structure and objects;
  - Processing user input in HTML Forms;
  - Introduction to HTML5.
- **My contributions**: I contributed with several HTML and CSS exercises for the students to complete in their TP lectures. The goal was to introduce the students to the development of web pages with HTML and CSS.

PD (Processamento de Dados) *2015–2016*

- **Degree**: Licenciatura em Biologia (2nd year)
- **My lectures**: TP – invited teacher
- **Topics**:
  - Introduction to data processing;
  - Introduction to Python (data types and data structures);
  - Introduction to Regular Expressions;
  - Introduction to Biomedical Web Services;
  - Database management systems.
- **My contributions**: I created all the practice class materials – a tutorial to guide students in their project, which was to process a large collection of protein and metabolic information extracted from widely-known biomedical web services, including protein sequences and metabolic pathway data. The goal was to teach students how to process biology-related data with a programming language and an underlying database.

Prog I (Programação I a outras licencituras) *2015–2016*

- **Degree**: Several B.Sc and M.Sc offered by Faculdade de Ciências da Universidade de Lisboa
- **My lectures**: TP
- **Topics**:
  - Computation: computability and Turing machines;
  - Algorithms: exhaustive search, approximation search and bisection search;
  - Programming methods: attribution and verification, decision, iteration and recursion, abstraction and specification, cloning;
  - Programming languages: expressions and types, precedence and associativity, functions, scope, libraries and modules;
  - Data structures: sequences, tuples, lists and dictionaries;
  - Files;
  - Software development: reading and writing, documentation, assertions and exceptions, test and debugging.
- **My contributions**: I contributed to the practice class materials by suggesting new exercises and modifications to existing ones.

Pedagogical surveys

Table 2: Pedagogical survey results. Students could either not answer each question or answer from 1 (strong disagreement) to 4 (strong agreement). Results for each question show the average for the students that answered the question.
Legend: *Q1*: Did the professor lecture with clarity? *Q2*: Did the professor answer questions with clarity? *Q3*: Was the professor available for outside-of-class contact & support? *Q4*: Was there a good pedagogical relation between professor and students? *Q5*: What is your global appreciation of the professor? *PL*: Prática de Laboratório (Practice lesson); *TP*: Teórico-Prática (Theoretical & Practical lesson); ∗: Course still ongoing, results unavailable.

	2015–2016					
	1st semester			2nd semester		
	IC		Prog I	ASW	ITW	
Question	TP	PL	TP	PL	TP	PL
Q1	3.86	3.86	3.69	∗	∗	∗
Q2	3.86	3.80	3.72	∗	∗	∗
Q3	3.70	3.71	3.62	∗	∗	∗
Q4	3.93	3.86	3.80	∗	∗	∗
Q5	3.87	3.88	3.79	∗	∗	∗

Teaching materials

ASW (Aplicações e Serviços na Web)

- I produced a guide to explain HTML forms, along with JavaScript access to the content of the forms and the content of the GET arguments passed with the URL. The goal was to teach students how to deal with user-input, as well as to introduce them to the idea of information flowing with the HTTP request from a form to the next web page. This material was made available to the students through moodle.

ITW (Introdução às Tecnologias da Web)

- I contributed with several HTML and CSS exercises for the students to complete in their TP lectures. The goal was to introduce the students to the development of web pages with HTML and CSS. This material was made available to the students through moodle.

PD (Processamento de Dados)

- I created all the practice class materials – a tutorial to guide students in their project, which was to process a large collection of protein and metabolic information extracted from widely-known biomedical web services, including protein sequences and metabolic pathway data. The goal was to teach students how to process biology-related data with a programming language and an underlying database. This material was made using google-docs and published as an HTML page, which was made available to the students through moodle.
- **URL**: `https://docs.google.com/document/d/1fhAJ3oL7Jx1W9m-zFs8ZxPpdFB59vEZHLWv9VGTOtDU/pub`

B.2 Jury and Examinations

Invitation to M.Sc in Bioinformatics and Computational Biology

2016

- **Student**: Samuel Viana
- **Supervisors**: Daniel Faria (Instituto Gulbenkian de Ciência) and Catia Pesquita (Universidade de Lisboa)
- I was formally invited by the supervisors; the examination is expected to happen in July.

B.3 Teaching-related activities

Invited participation in courses

AW (Aplicações na Web) *2014–2015*

- **Degree**: Mestrado em Engenharia Informática
- I presented a lecture on Semantic Web (SW) with the following topics:
  - The problem of ambiguity that SW tries to solve;
  - Rule-based inference;
  - RDF statements;
  - Several of the SW languages (RDF, OWL, SPARQL);
  - Objects *vs.* Classes *vs.* Instances
  - An introduction to several of the layers of the SW, including URIs, XML, RDF, Ontologies and Rules;
  - Real-world examples of SW in action: Semantic wikis, FOAF project, RDFa, hCalendar, Linked Data Project.

Bioinformatics & Computational Modelling *2013–2014*

- **Degree**: PhD Program in Biological Systems – Functional & Integrative Genomics
- I presented a practical lecture on Bioinformatics, specifically on the use of Python in biomedical data processing. This included:
  - Basic python datatypes and functions;
  - Introduction to BioPython, a package with access to several functions dedicated (i) to biological data processing and (ii) to widely-known biomedical web services;
  - Exercises directed at learning the inners of BioPython, specifically to process protein sequences (using web services to access the SwissProt database and the BLAST algorithm)
  - Introduction to the Gene Ontology and Semantic Similarity

Ontologias Aplicadas às Ciências *2013–2014*

- **Degree**: M.Sc. class available to several M.Sc. students at FCUL
- I collaborated on the practical classes by designing a project and supervising a group of students in implementing the project's specifications. The projects were designed so that students would develop an intuition about ontology development, ontology matching, and semantic similarity. Two of these projects resulted in publications at national and international level.

Bioinformática *2011–2012*

- **Degree**: B.Sc. class available to several B.Sc. students at FCUL
- I was a teaching assistant on the practical classes by invitation of the course's head professor. I created a project specification based on ontology development and text-mining, and supervised the students in their implementations.

C University Mission

C.1 Participation in socially relevant projects

FCUL Open Day

2015

- I helped devise a game-based activity that used two Bioinformatics tools developed at LASIGE to show students students the current challenges in exploring textual data in the biomedical domain and how semantic-based tools can assist the research process.

Semana da Ciência em Movimento

2014

- I helped create and direct the BioIn4matix Run, a peddypaper activity that took 50 secondary school students throughout a physical journey around the FCUL Campus as well as a conceptual journey several around Bioinformatics research steps.

C.2 Scientifically relevant projects and organizations of scientific, professional or cultural interest

2as Jornadas GIMGAS (Grupo de Internos de Medicina Geral e Familiar de Almada-Seixal)

2015

- This is a conference organized by internal General Practitioner residents from Almada and Seixal, open to interns and recently specialized doctors. I was responsible for creating a registration website for the conference, including managing the registration for the several workshops presented at the conference. The conference was attended by more than 100 doctors.

D Ongoing Research

My current research efforts are focus on the application of semantic-web related technologies to multi-disciplinary data, especially biomedical data, which usually exhibits high degree of multidisciplinarity. I summarize my efforts in four topics.

D.1 Multidisciplinary data-mining

Application of information-mining technologies to multidisciplinary data, including text- and data-mining of biomedical related information, with a particular emphasis on the use of semantic similarity measures to that effect. Data sources include health records, metabolic pathway descriptions, scientific publications, digital representations of biomedical entities, *etc*.

D.2 Machine-learning in biomedical data

Application of machine-learning algorithms (in particular neural networks and deep-learning algorithms) to biomedical data. I have not yet started to research this topic, but I will in the very near future.

D.3 Semantic similarity

Exploration of several ideas to improve semantic similarity measures in a multidisciplinary content. For the moment, this consists in a set of measures of relevance of concepts in an ontology.

D.4 Data quality

Assessment of data quality by measuring some characteristics of the associated metadata, which including measuring the specificity and coverage of the metadata.

E Full publication list

E.1 Journal papers

[1] Andre Lamurias, **João D. Ferreira**, and Francisco M. Couto. "Improving chemical entity recognition through h-index based semantic similarity." In: *Journal of Cheminformatics* 7.Suppl 1 Text mining for chemistry and the CHEMDNER track (Jan. 2015), S13. ISSN: 1758-2946. DOI: 10.1186/1758-2946-7-S1-S13.

[2] Andre Lamurias, **João D. Ferreira**, and Francisco M. Couto. "Identifying interactions between chemical entities in biomedical text". In: *Journal of Interactive Bioinformatics (JIB)* 11.3 (2014), pp. 1–18. ISSN: 1613-4516. DOI: 10.2390/biecoll-jib-2014-247.

[3] Catia Pesquita, **João D. Ferreira**, Francisco M. Couto, and Mário J. Silva. "The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources." In: *Journal of Biomedical Semantics* 5.1 (2014), p. 4. ISSN: 2041-1480. DOI: 10.1186/2041-1480-5-4.

[4] **João D. Ferreira**, Janna Hastings, and Francisco M. Couto. "Exploiting disjointness axioms to improve semantic similarity measures." In: *Bioinformatics* 29.21 (2013), pp. 2781–2787. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt491.

[5] **João D. Ferreira**, Daniela Paolotti, Francisco M. Couto, and Mário J. Silva. "On the usefulness of ontologies in epidemiology research and practice." In: *Journal of epidemiology and community health* (Nov. 2012), pp. –3. ISSN: 1470-2738. DOI: 10.1136/jech-2012-201142.

[6] **João D. Ferreira** and Francisco M. Couto. "Semantic Similarity for Automatic Classification of Chemical Compounds". In: *PLoS Computational Biology* 6.9 (Sept. 2010). Ed. by John B. O. Mitchell, e1000937. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000937.

E.2 Conferences (oral presentations & posters)

[7] **João D. Ferreira**, Bernard de Bono, and Francisco M. Couto. "From data to knowledge: a tool for clustering multi-scale resources for physiology research". In: *Virtual Phisiology Human Conference*. 2014.

[8] Andre Lamurias, **João D. Ferreira**, and Francisco M. Couto. "Chemical named entity recognition: Improving recall using a comprehensive list of lexical features". In: *8th International Conference on Practical Applications of Computational Biology & Bioinformatics*. Vol. 294. AISC. 2014, pp. 253–260. DOI: 10.1007/978-3-319-07581-5_30.

[9] **João D. Ferreira**, Catia Pesquita, Mário J. Silva, and Francisco M. Couto. "Digital preservation of epidemic resources: coupling metadata and ontologies". In: *International Conference on Preservation of Digital Objects*. Ed. by José Borbinha, Michael Nelson, and Steve Knight. 2013. URL: http://xldb.di.fc.ul.pt/xldb/publications/Ferreira.etal:DigitalPreservationOf:2013_document.pdf.

[10] Catia Pesquita, **João D. Ferreira**, Francisco M. Couto, and Mário J. Silva. "Semi-automated annotation of epidemiological resources". In: *Epiwork International Workshop on Digital Epidemiology*. Epiwork International Workshop on Digital Epidemiology, 2013. URL: http://webpages.fc.ul.pt/~fjcouto/files/abstract%20cpesquita-epiwork2013.pdf.

[11] David S. Batista, **João D. Ferreira**, Francisco M. Couto, and Mário J. Silva. "Toponym Disambiguation using Ontology-based Semantic Similarity". In: *Computational Processing of the Portuguese Language*. Vol. 7243. 2012, pp. 179–185. ISBN: 9783642288845. DOI: 10.1007/978-3-642-28885-2_20.

[12] Francisco M. Couto, **João D. Ferreira**, João Zamite, Carlos Santos, Tiago Posse, Paulo Graça, Dulce Domingos, and Mário J. Silva. "The Epidemic Marketplace Platform: towards semantic characterization of epidemiological resources using biomedical ontologies". In: *International Conference on Biomedical Ontologies*. 2012. URL: http://ceur-ws.org/Vol-897/demo_1.pdf.

[13] **João D. Ferreira**, Catia Pesquita, Francisco M. Couto, and Mário J. Silva. "Bringing epidemiology into the Semantic Web". In: *International Conference on Biomedical Ontologies*. Vol. 897. 2012. URL: http://ceur-ws.org/Vol-897/session1-paper02.pdf.

[14] **João D. Ferreira** and Francisco M. Couto. “Generic semantic relatedness measure for biomedical ontologies”. In: *International Conference on Biomedical Ontologies*. Vol. 833. 2011, pp. 117–123. URL: http://ceur-ws.org/Vol-833/paper16.pdf.

[15] Bruno Tavares, Hugo Bastos, Daniel Faria, **João D. Ferreira**, Tiago Grego, Catia Pesquita, and Francisco M. Couto. “The Biomedical Ontology Applications (BOA) framework”. In: *International Conference on Biomedical Ontologies*. Vol. 833. 2011, pp. 300–301. URL: http://ceur-ws.org/Vol-833/paper56.pdf.

E.3 Thesis (see section F)

[16] **João D. Ferreira**. “Semantic Similarity Across Biomedical Ontologies”. PhD Thesis. Universidade de Lisboa, 2016.

[17] **João D. Ferreira**. “Structural and semantic similarity metrics for chemical compound classification”. Master’s Dissertation. Universidade de Lisboa, 2010.

E.4 Other

[18] **João D. Ferreira**, Catia Pesquita, Francisco M. Couto, and Mário J. Silva. *Epiwork Deliverable 3.5: Epidemic Data Ontology*. Tech. rep. University of Lisbon, Faculty of Sciences, LASIGE, Jan. 2012.

[19] **João D. Ferreira**, David S. Batista, Francisco M. Couto, and Mário J. Silva. *The Geo-Net-PT / Yahoo ! GeoPlanet TM concordance*. Tech. rep. October. Departamento de Informática – Faculdade de Ciências - Universidade de Lisboa, 2010. DOI: DOI:10455/6677.

[20] Tiago Grego, **João D. Ferreira**, Catia Pesquita, Hugo Bastos, Diogo Vila Viçosa, João M. Freire, and Francisco M. Couto. *Chemical and Metabolic Pathway Semantic Similarity*. Tech. rep. Department of Informatics, Faculty of Sciences, University of Lisbon, 2010. DOI: DOI:10455/3335.

F Education

Ph.D. in Computer Science – Bioinformatics

Dec. 2010 – Jan. 2016

Universidade de Lisboa

- **Thesis**: Semantic Similarity Across Biomedical Ontologies
- **Supervisor**: Prof. Dr Francisco M. Couto
- **Grade**: Aprovado com Distinção e Louvor
- **Abstract**: The need to compare complex entities is relevant in all the areas of science. In medicine, for example, comparing a clinical case to a database of previous cases can be extremely helpful when trying to diagnose a disease or deciding the most appropriate treatment for a patient.

  Recent developments in knowledge representation, in particular the creation of the Web Ontology Language (OWL), have lead to a rise in the amount of knowledge that is being stored in *ontologies*, which represent, in machine-readable format, the known facts about reality. With the help of ontologies, statements like "**Influenza** is an **Infectious disease**" can be processed by computers, which, in turn, can be used to create new knowledge. In particular, *semantic similarity* has emerged to explore these ontologies as a way to compare entities annotated with the ontology concepts.

  Semantic similarity has been extensively studied in the last decade, but some problems still persist. While there are algorithms to compare entities annotated with concepts from the same ontology, the possible ways to use *more than one ontology* are still in an early phase of study. For example, comparing a metabolic pathway using both the associated molecular functions and the metabolites converted in the pathway should, in principle, yield a higher precision than would be achieved with methodologies that rely on either one of the two domains independently. Comparing concepts from *different domains* and entities annotated with concepts from different domains is yet an unexplored area, but necessary to tackle multidisciplinary biomedical resources, *e.g.* to compare two clinical cases, the relationships between symptoms, diseases, blood screening results, *etc.* should provide a more insightful and precise value of similarity.

  In this document, I explain the basic concepts needed to understand the problem of semantic similarity, how it is being solved, and how I propose to extend this notion so that it can be applied to more than one ontology and, more significantly, to more than one domain of knowledge.

M.Sc. in Biochemistry

Sep. 2008 – June 2010

Universidade de Lisboa

- **Thesis**: Structural and semantic similarity metrics for chemical compound similarity
- **Supervisor**: Prof. Dr Francisco M. Couto
- **Grade**: 19 out of 20
- **Abstract**: Over the last few decades, there has been an increasing number of attempts at creating systems capable of comparing and classifying chemical compounds based on their structure and/or physicochemical properties. While the rate of success of these approaches has been increasing, particularly with the introduction of new and ever more sophisticated methods of machine learning, there is still room for improvement. One of the problems of these methods is that they fail to consider that similar molecules may have different roles in nature, or, to a lesser extend, that disparate molecules may have similar roles.

  This thesis proposes the exploitation of the semantic properties of chemical compounds, as described in the CHEBI ontology, to create an efficient system able to automatically deal with the binary classification of chemical compounds. To that effect, I developed Chym (Chemical Hybrid Metric) as a tool that integrates structural and semantic information in a unique hybrid metric.

  The work here presented shows substantial evidence supporting the effectiveness of Chym since it has outperformed all the models with which it was compared. Particularly, it achieved accuracy values of 90.9%, 87.7% and 84.2% when solving three classification problems which, previously, had only been solved with accuracy values of 81.5%, 80.6% and 82.8% respectively. Other results show that the tool is appropriate to use even if the problem at hand is not well represented in the CHEBI ontology. Thus, Chym shows that considering the semantic properties of a compound helps solving classification problems.

Therefore, Chym can be used in projects that require the classification and/or the comparison of chemical compounds, such as the study of the evolution of metabolic pathways, drug discovery or in preliminary toxicity analysis.

B.Sc. in Biochemistry

Sep. 2005 – June 2008

Universidade de Lisboa

- **Discrimination of some grades**:
  - Análise e Tratamento de Dados em Bioquímica *[Data Analysis and Processing in Biochemistry]*: 19
  - Bioquímica Computacional *[Computational Biochemistry]*: 19
  - Simulação Computacional *[Computacional Simulation]*: 20
- **Final grade**: 18 out of 20