MERCÈ CROSAS, PH.D.

UNIVERSITY RESEARCH DATA MANAGEMENT OFFICER
Harvard University

CHIEF DATA SCIENCE AND TECHNOLOGY OFFICER
Institute for Quantitative Social Science, Harvard University

mcrosas@g.harvard.edu http://scholar.harvard.edu/mercecrosas @mercecrosas

EDUCATION

1994 - 1997	**Post-Doctoral Fellow, Submillimeter Array, Radioastronomy Division,** advisor: Karl Menten; Harvard-Smithsonian Center for Astrophysics, Cambridge, MA
1992 - 1994	**Pre-Doctoral Fellow, Institute for Atomic and Molecular Physics**, advisor: Alex Dalgarno; Harvard University, Cambridge, MA
1989 - 1994	**Ph.D. in Astrophysics**, advisor: John Weisheit; Rice University, Houston, TX
1984 - 1989	**B.S. in Physics**; Universitat de Barcelona, Barcelona, Spain

CAREER HISTORY

HARVARD UNIVERSITY RESEARCH DATA MANAGEMENT OFFICER

Harvard University Information Technology, Harvard University
2019 – Present

Provide leadership to mature Harvard's data management and governance practices. Work in close collaboration with key constituencies in Research, Information Technology, and the Library to evolve Harvard's research data life cycle management framework and guide university policy, process, and procedures for research data. Dr. Crosas brings to this role a wealth of experience in data management architecture and international community standards as well as the vision to make data more accessible while preserving privacy.

CHIEF DATA SCIENCE AND TECHNOLOGY OFFICER

Institute for Quantitative Social Science, Harvard University,
2004 – Present (current title since October 2015)

Lead the vision and strategic direction of data sharing and data analysis projects developed at the Institute for Quantitative Social Science at Harvard University. Evaluate research and academic needs, and provide high-level requirements and priorities for the software projects, including the Dataverse (data repository software), OpenDP (differential privacy), and Consilience (text analysis). Manage existing research collaborations and grants, and pursue new

ones. Provide advice and develop ideas for new software projects and technologies. Supervise the data curation and data management team, usability and user experience team, and the data science services team. Participate in outreach nationally and internationally, and contribute to standards and best practices on topics related to research data.

DIRECTOR OF SOFTWARE DEVELOPMENT/IT

Cantata Laboratories, Cambridge, MA
2002 -2004
Led the software development of the Lab Information Management system for Mass Spectrometry experimental data and analysis results in a biotechnology startup. Managed budget, roadmaps and resource allocations.

MANAGER OF SOFTWARE DEVELOPMENT

Cereon Genomics, Cambridge, MA
2001 -2002
Led the software development of the Data Management System for SNP discovery and genotyping in a biotechnology startup. Worked closely with scientist, bioinformaticians and software engineers to design, architect and implement the software system and database.

SENIOR MANAGER OF SOFTWARE DEVELOPMENT

WebCT, Peabody, MA
1999 -2001
Led the software development of the WebCT Learning Management System, working closely with user experience experts, designers, education researchers and software engineers. In 2001, served in the software architecture team as a senior architect of the entire learning management system.

ASTROPHYSICIST, SOFTWARE ENGINEER

Harvard-Smithsonian Center for Astrophysics
1997 -1999
Built a two-dimensional Monte Carlo simulation for radiative transfer and molecular emissions in circumstellar envelopes, to explain radio spectral line observations from these old stars. As a member of the Submillimeter Array (SMA) project, implemented software for building and configuring the SMA radio interferometer located at Mauna Kea, Hawaii.

SOFTWARE

Dataverse (co-PI)

The Dataverse is a software platform for building data repositories to share, archive and cite research data sets. It is used world-wide with 67 deployed data repositories in 6 continents, and with an active international open-source community. (Technologies and languages: Java, Javascript, Glassfish, PostgreSQL, Solr/Lucene, Python)

OpenDP (co-PI)

OpenDP is a community effort to build trustworthy, open-source software tools for statistical analysis of sensitive private data. The tools offered by OpenDP provide the rigorous protections of differential privacy for the individuals who may be represented in confidential data and statistically valid methods of analysis for researchers who study the data.

DataTags (co-PI)

The DataTags system provides a platform for assigning a machine-actionable policy to a sensitive data set, which describes its security and access requirements. (Technologies and languages: Java, Javascript, Scala, Glassfish)

Consilience (supervisor)

The Consilience software provides a platform to analyse a large number of text documents and find the most suitable clustering solution. (Technologies and languages: Java, D3.js, Spark, R, MongoDB, Solr, Glassfish)

RESEARCH GRANTS

2020 - present	OpenDP, **Alfred P. Sloan Foundation** , co-PIs: Salil Vadhan, Gary King, James Honaker, Mercè Crosas
2020 - present	EAGER, Sharing Knowledge, Building community, Introducing a Journal Editors' Discussion Interface (JEDI), **National Science Foundation** , co-PIs: Colin Elman, Diana Kapiszewski, Margaret Levenstein, Thu-Mai Christian, Mercè Crosas
2018 - 2020	A proposed quantitative index to understand and evaluate the health of open source projects, **Alfred P. Sloan Foundation** , PI: Mercè Crosas
2018 - 2019	Understanding What Constitutes a Vibrant Open-Source Community, **Institute of Museum and Library Services** , PI: Mercè Crosas
2018 - 2020	Increasing Scientific Dataset Quality Through Reproducibility and Curation Tools and Targeted Services in Dataverse Repositories, **Alfred P. Sloan Foundation** , PI: Mercè Crosas
2015 - 2019	Applying Theoretical Advances in Privacy in Computational Social Science , **Alfred P. Sloan Foundation** , Co-PIs: Micah Altman, Salil Vadhan, Gary King, Mercè Crosas
2015 - 2019	Piloting functionality of a biomedical research data management system with structural biology datasets, **Leona M. and Harry B. Helmsley Charitable Trust** , Co-PIs: Piotrek Sliz, Mercè Crosas

2015 - 2019	Towards a FAIR Digital Ecosystem in the Cloud, **National Institutes of Health**, , Co-PIs: Mercè Crosas, Tim Clark, Martin Fenner
2012- 2018	TWC: Frontier: Privacy for Social Science Research, **National Science Foundation**, PI: Salil Vadhan (Crosas is a Co-Investigator)
2015 - 2017	A Bridge from Publishing Words to Publishing Data **Alfred P. Sloan Foundation**, Co-PIs: Gary King, Mercè Crosas
2016 - 2017	Preparing Social Science Research for the Potential Inversion of Its Largest Successes and Failures, **Alfred P. Sloan Foundation**, Co-PIs: Gary King, Mercè Crosas
2015 - 2016	Citation++: Data Citation, Provenance and Documentation **National Science Foundation**, Co-PIs: Margo Seltzer, Mercè Crosas, Gary King
2012 - 2016	BIGDATA: Mid-Scale: ESCE: DCM: Collaborative Research: DataBridge – A Sociometric System for Long-tail Science Data Collections, **National Science Foundation** , PI: Arcot Rajasekar (Crosas is a Co-Investigator)
2013 - 2015	Collaborative Research: Center for Historical Information and Analysis **National Science Foundation**, PI: Patrick Manning (Crosas is a Co-Investigator)

SELECTED COMMITTEES, BOARDS, WORKING GROUPS

- Expert Group **Catalunya 2022**
- Advisory Board of Harvard Data Science Review
- Expert Committee for the National Academies of Science, Engineering, and Medicine consensus report on Realizing Opportunities for Advanced and Automated Workflows in Scientific Research
- Services and Technology Steering Group of DataCite
- Advisory Board of Authorea
- Advisory Board OpenAIRE

- Advisory Board of LYRASIS project on Open Source Software
- Editorial Board of Scientific Data by Nature Publishing Group
- Steering Committee of Data-PASS (Preservation Alliance for the Social Sciences)
- Steering Committee of the National Data Service
- Advisory Board of the Qualitative Data Repository
- Advisory Board of CRADLE (Curating Research Assets and Data using LifeCycle Education)
- Advisory Board for the Databrary project (sharing annotating research videos)
- Advisory Board of Force11 (improve scholarly communication and research)
- Co-Chair of Data Citation Implementation group
- Member of NIH B2DK BioCaddie Community Engagement Working group
- Member of Research Data Alliance (RDA) Data Publishing Interest group
- Member of Seamless Astronomy Group, Harvard-Smithsonian Center for Astrophysics
- Member of the international Software Citation Group
- IEEE Big Data Meeting 2015, Organizing Committee
- IACS Symposium 2015, Privacy in a Networked World, Organizing Committee
- National Data Service Consortium 2016, Organizing Committee
- FORCE2016, Research Communication and eScholarship Conference, Organizing Committee (2017)

RECENT PUBLICATIONS

Repository Approaches to Improving the Quality of Shared Data and Code. MDPI Data. 2021

Buckee C, Balsari S, Chan J, **Crosas M**, Dominici F, Gasser U, Grad Y, Grenfell B, Halloran ME, Kraemer M, et al. Aggregated mobility data could help fight COVID-19. Science. 2020;23 March (Letters)

Alexander S, Jones K, Bennet N, Buden A, Cox M, **Crosas M,** Game E, Geary J, Hardy D, Johnson J, et al. Qualitative data sharing and synthesis for sustainability science. Nature Sustainability. 2020;(3) :81-88

Jacobsen A, de Azevedo RM, Juty N, Batista D, Coles S, Cornet R, Courtot M, **Crosas M**, Dumontier M, et al. FAIR principles: Interpretations and implementation considerations. Data Intelligence

Wilkinson MD, Dumontier M, Sansone S-A, Olavo L, Prieto M, Batista D, McQuilton P, Kuhn T, Rocca-Serra P, **Crosas M**, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. Nature-Springer Scientific Data. 2019;6 (174)

Fenner M, **Crosas M**, Grethe J, Kennedy D, Hermjakob H, Rocca-Serra P, Durand G, Berjon R, Karcher S, Martone M, et al. A Data Citation Roadmap for Scholarly Data Repositories. Nature-Springer Scientific Data. 2019;6 (28)

Crosas M, Gautier J, Karcher S, Kirilova D, Otalora G, Schwartz A. Data policies of highly-ranked social science journals. SocArXiv. 2018; March.

Pasquier T, Lau M, Trisovic A, Boose E, Couturierer B, **Crosas M**, Ellison A, Gibson V, Jones C, Seltzer M. If These Data Could Talk. Nature Scientific Data. 2017

Pasquier, T., Lau, M., Trisovic, A., Boose, E., Coutirierer, B., **Crosas M**., Ellison, A., Gibson, V., Jones, C., Seltzer, M. 2016, "If These Data Could Talk" Nature Scientific Data

Bierer. B, **Crosas M**., Pierce, H. 2017, "Data Authorship as an Incentive to Data Sharing" New England Journal of Medicine, Sounding Board

McKinney, B., Meyer. P., **Crosas M**., Sliz, P. 2016, "Extension of Research Data Repository System to Support Direct Compute Access to Biomedical Datasets" The Annals of the New York Academy of Science

Bar-Sinai M, Sweeney L, **Crosas M**. 2016, "Data Handling Policy Spaces and the Tags Language" Proceedings of the International Workshop on Privacy Engineering. IEEE

Peter A. Meyer, Stephanie Socias, Jason Key, Elizabeth Ransey, Emily C. Tjon, Alejandro Buschiazzo, Ming Lei, Chris Botka, James Withrow, David Neau, Kanagalaghatta Rajashankar, Karen S. Anderson, Richard H. Baxter, Stephen C. Blacklow, Titus J. Boggon, Alexandre M.J.J. Bonvin, Dominika Borek, Tom J. Brett, Amedeo Caflisch, Chung-I Chang, Walter J. Chazin, Kevin D. Corbett, Michael S. Cosgrove, Sean Crosson, Sirano Dhe-Paganon, Enrico Di Cera, Catherine L. Drennan, Michael J. Eck, Brandt F. Eichman, Qing R. Fan, Adrian R. Ferre´-D'Amare´, J. Christopher Fromme, K. Christopher Garcia, Rachelle Gaudet, Peng Gong, Stephen C. Harrison, Ekaterina E. Heldwein, Zongchao Jia, Robert J. Keenan, Andrew C. Kruse , Marc Kvansakul, Jason S. McLellan, Yorgo Modis, Yunsun Nam, Zbyszek Otwinowski1, Emil F. Pai, Pedro Jose´ Barbosa Pereira, Carlo Petosa, C.S. Raman, Tom A. Rapoport, Antonina Roll-Mecak, Michael K. Rosen47, Gabby Rudenko, Joseph Schlessinger, Thomas U. Schwartz, Yousif Shamoo, Holger Sonderman, Yizhi J. Tao, Niraj H. Tolia, Oleg V. Tsodikov, Kenneth D. Westover, Hao Wu, Ian Foster, James S. Fraser, Filipe R.N.C. Maia, Tamir Gonen, Tom Kirchhausen, Kay Diederichs, **Mercè Crosas** & Piotr Sliz. 2016, "Data publication with the structural biology data grid supports live analysis" *Nature Communications* 7, Article number: 10882

Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Olavo Bonino da Silva Santos, Philip Bourne, Jildau Bouwman, Anthony Brookes, Tim Clark, **Mercè Crosas**, Ingrid Dillo,

Olivier Dumon, Scott Edmunds, Chris Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair Gray, Paul Groth, Carole Goble, Jeffrey Grethe, Jaap Heringa, PETER t Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott Lusher, Maryann Martone, Albert Mons, Abel Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons., 2016. "The FAIR Guiding Principles for scientific data management and stewardship" Scientific Data,

Sweeney L, **Crosas M**, Bar-Sinai M. 2015, "Sharing Sensitive Data with Confidence: the DataTags System" Journal of Technology Science

Altman M, Castro E, **Crosas M**, Durbin P, Garnett A, Whitney J. 2015, "Open Journal Systems and Dataverse Integration-- Helping Journals to Upgrade Data Publication for Reusable Research" Code4Lib Journal, Issue 30

Starr J, Castro E, **Crosas M**, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. 2015, "Achieving human and machine accessibility of cited data in scholarly publications" PeerJ Computer Science, 1:e1 https://dx.doi.org/10.7717/peerj-cs.1

Goodman, A., Pepe, A., Blocker, A.W., Borgman, C.L., Cranmer,K., **Crosas, M.,** Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D.W., Kashyap,V., Mahabal, A., Siemiginowska, A., Slavkovic, A., 2014. 10 Simple Rules for the Care and Feeding of Scientific Data, PLoS Comput Biol, doi:10.1371/journal.pcbi.1003542

Pepe, A., Goodman, A., Muench, A., **Crosas, M.,** Erdmann, C., 2014. How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. PLoS ONE, DOI: 10.1371/journal.pone.0104798

Crosas, M., 2013. A Data Sharing Story, Journal of eScience Librarianship, 2013, 1(3), 173-179, http://dx.doi.org/10.7191/jeslib.2012.1020

Altman, M., **Crosas, M**. 2013. The Evolution of Data Citation: From Principles to Implementation, IASSIST Quarterly, p. 62

Crosas, M, 2011. The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data. D-Lib Magazine, volume 17.

TEACHING AND PRESENTATIONS

- Teaching and mentoring have been essential parts of my work in both industry and academia. In my leadership role building software solutions for research, I have mentored engineers and data scientists one-on-one, as well as provided training, tutorials and talks on how to use new technology frameworks, understand a system's architecture, object and data models and more.
- Within Harvard, I have also lectured and mentored in research design, data management, data security, and research reproducibility to a wider audience of undergraduate and graduate students, post-doctoral fellows, researchers and staff.
- Outside Harvard, I have participated in numerous conferences, symposia, and workshops, either giving lectures or contributing to a panel discussion to educate on data management or data science related topics, or to provide guidance on building new data technologies, standards and protocols to improve research.

Below is a list of selected activities that reflect a diverse teaching, lecturing, and mentoring experience.

Internships at IQSS, 2011 – present

- Serve as mentor Research Experience Undergraduates (REUs) working on our NSF Privacy Tools project (2014, 2015, 2016). Develop and teach orientation sessions and tutorials on Dataverse architecture and APIs, as well as on the DataTags project, with a focus on codifying privacy regulations to share sensitive data through open repositories.
- Start and run a Data Science internship program at IQSS (2012, present): includes mentoring and teaching tutorials on Dataverse architecture, on building data exploration tools, and text analysis tools. The first summers, the program included one or two interns per summer. On 2015 there were a total of 13 interns in this program.
- Mentor summer student interns to implement academic websites. (2011)

Digital Problem Solving mentorship (Berkman Center), 2014

- Mentor (with postdoctoral fellow Vito D'Orazio) a group of five Harvard undergraduates to design and build a generic data exploration and analysis tools, retrieving data from repositories.

Lunch Technology Talks, IQSS, 2011 – Present

- Informal lectures (2 to 3 per year) for an audience of software engineers, user-interface developers and information scientists on new web application technologies, data systems and software frameworks.

Training Sessions, IQSS, 2011 - Present

- Led monthly and quarterly training session for academic web site building, and preparing, managing and sharing research data

Mentorship and leadership of software teams (WebCT, Cereon Genomics, 1999-present)

- As part of leading software development teams in the 16 years, teach and mentor software engineers on object and data model design, systems architecture, business logic, user experience and user interface frameworks, and cross-disciplinary team work.

Harvard's GSAS Guidance

- Over the last five years, contributed occasionally to tutorials for new Harvard graduate students to guide them on working with data, using the appropriate research tools for collecting and analyzing data, and reproducing their own research.

SELECTED SEMINARS, LECTURES, AND WORKSHOPS:

Recent Keynotes:

- "Research Data Management with Dataverse" Haverford College, 2021
- "Enhancing Collaboration and Access to Data throughout the Research Lifecycle" Sanford University, 2021
- Cicle de Conferències sobre la gestó de dades de recerca (Research Data Management), organitzat pel CSUC
    - "El Data Commons o com facilitar la col·laboració i accés a les dades de recerca; una visió"
    - "La gestió FAIR de les dades de la recerca amb Dataverse",
    - "Serveis de dades i computació per al suport del cicle de recerca a una universitat"
- "Data sharing with Dataverse" National Library of Medicine, 2020
- "FAIR and Responsible Data Sharing with Dataverse". Beilstein Open Science Symposium, 2020
- "Dataverse and OpenDP: Tools for privacy-protecting analysis in the cloud" Red Hat Research Day, 2020
- "Harvard Data Commons" Tromso, Norway, 2020
- "Recommendations for Implementing Open Science" EPFL, Lausanne, Switzerland, 2019
- "OECD workshop on the revision of the recommendation concerning access to research data from public funding", OECD, Paris, 2010
- "Research Data Management at Harvard University, Data Sharing, and the Dataverse Project" Portugal, 2018
- "Dataverse", Science Gateways, Austion, Texas
- "Data Sharinf for Better Science" Max-Planck Institute for Radioastronomy, Bonn, Germany, 2017
- "Data Sharing for Better Science and Better ", XVII Congreso SESPAS, Barcelona, Spain, 2017

Lectures for Harvard FAS Responsible Conduct of Research, 2013 - 2020

- Lecture on Research Data Management and Security, winter 2016 (co-led with Ara Tahmassian, Office of Vice Provost of Research).

- Lecture on Data Acquisition and Data Management, summer 2015 (co-led with Kristen Bolt, Office of Vice Provost for Research).
- Lecture on Research Data Retention, Fall, 2013

CRADLE Research Data Management & Sharing MOOC, Advisory Board, 2013-2016

- Contribute to the creation of a MOOC on data management and sharing, led by the University of North Caroline Chapel Hill and the Institute of Museum and Library Services.
- Member of the advisory board to help build the syllabus and review course materials.

FORCE2016 Conference Session on Communicating Science, April, 2016

- Led a session on communicating science, including lectures and a panel discussion and bringing from different fields: Steven Pinker (Harvard), César Hidalgo (MIT), Christie Nicholson (Alan Alda Institute).
- Presentation on the history of communicating science, from 1665 to present, from print to digital, from verbal text to data.

Data Citation Implementation Workshop, February 2016

- Presentation on how to implement data citation in a data repository.

Seminar for the National Information Standards Organization, December 2015

- On Addressing the New Challenges of Data Sharing: Large-Scale Data and Sensitive Data.

Harvard's Technology Science Seminar, December 2015

- On Sharing Sensitive Data with Confidence, co-led with Latanya Sweeney and Michael Bar-Sinai.

Massachusetts Open Cloud Workshop, November, 2015

- Presentation and panel discussion on integrating a data repository (Dataverse) with the new Massachusetts Open Cloud (MOC), led by Boston University's Hariri Institute for Computing.

Workshop on The Future of the Commons: Data, Software and Beyond, November 2015

- Presentation and panel discussion on the Data Commons in a workshop organized by CENDI, NFAIS and Research Data Alliance.

Harvard Research Access and Innovation Transparency Symposium, October 2015

- Lecture on Data Publishing in a symposium organized by Harvard Medical School.

IEEE HPEC Conference, September 2015

- Presentation and panel discussion on the Future Directions of Research Computing and Cyberinfrastucture , towards a Northeastern Research Computing Center.

Harvard-Purdue Data Management Symposium, June 2015

- Presentation and panel discussion on a data management, in a symposium organized by the Harvard Library and Purdue University Libraries.

Society for Scholarly Publishing Conference, June 2015

- Presentation in a session for journals editors on helping them establish data access and research transparency practices.

Workshop on Transparency and Reproducibility in Federal Evaluations, April 2015

- Presentations and panel discussions on software and data sharing solutions and on sharing sensitive data.

Harvard's Library Seminar, January 2015

- Lecture Data Management and Sharing (and the Dataverse repository) organized by Harvard Library.

Research Data Alliance (RDA) Plenary, September 2014

- Presentation on Archiving Social Science Data.

IRODS User Meeting, July 2014

- Tutorial on Sharing Data you Can't Share, co-led with Michael Bar-Sinai.

IASSIST Conference, June 2014

- Lecture on Data Publishing while Preserving Data Privacy.

Berkeley Institute for Transparency in Social Science (BITSS) Summer Institute, June 2014

- Lecture on Research Transparency Through Data Sharing.

Harvard's Sociology Department Course, 2013

- Lecture on Research Design (with Sociology Professor Jocelyn Viterna).

Harvard's IACS Seminars Series, October 2013

- Lecture on "The Care and Feeding of Scientific Data".

Harvard IT Summit

- Lecture on Collaboration in Science and Technology (2015).
- Data Sharing sessions (2012, 2013).

A list of my **publications** and **presentations** can also be found at:
http://scholar.harvard.edu/mercecrosas/publications
http://scholar.harvard.edu/mercecrosas/presentations