Availability of raw data in Spanish cancer research

Rut Lucas-Domínguez¹,²,³, Rubén Fernández¹, Yiming Liu¹,², Andrea Sixto-Costoya¹,²,⁴, Cristina Rius¹,² and Juan Carlos Valderrama-Zurián¹,²

Rut.Lucas@uv.es (0000-0002-4697-7832), rufergas@alumni.uv.es (0000-0003-0440-1330), Yiming.Liu@uv.es (0000-0006-3592-0175), Andrea.Sixto@uv.es (0000-0001-9162-8992), Cristina.Rius@uv.es (0000-0002-0572-5620), Juan.Valderrama@uv.es (0000-0001-5787-6853)

¹UISYS Group. Department of History of Science and Information Science. Faculty of Medicine and Dentistry. University of Valencia. Spain
²Unit associated with the Interuniversity Institute for Advanced Research on the Evaluation of Science and the University (INAECU) UC3M-UAM. Spain
³CIBERONC. Spain
⁴University Institute of Pure and Applied Mathematics (IUMPA). Valencia Polytechnic University. Spain

Abstract
This study aims to analyze the availability of research data deposited by Spanish authors in the area of cancer. A bibliographic search of raw data signed by at least one Spanish author and published in the period 2011-2021 was carried out using the Data Citation Index. This study revealed a total of 2,693 registers from 2,679 authors, being Gene Expression Omnibus, the main repository used.

1. Introduction
Around 10 million people died from cancer worldwide in 2020, of which approximately 4.5 million were women. In Europe, cancer is the second leading cause of death, almost equalling cardiovascular diseases as the first cause, with an estimated 30.4 million cancer cases and a mortality of 16.3 million people in 2040 (Sung et al., 2021). The COVID-19 pandemic has been an example showing that in addition to research and publication, it is essential to share research data to ensure scientific progress, reproducibility, and transparency (Lucas-Dominguez et al., 2021). The aim of this work has been to evaluate the availability and type of datasets produced in the area of oncology, establishing which are the main authors from Spanish institutions in scientific research in cancer.

2. Methods
A bibliographic search was performed in the Data Citation Index (DCI) of deposited data on cancer between 2011-2021, signed by at least one Spanish author, to obtain an overview in terms of temporal evolution, thematic categories, repositories used, authorships, institutions, and countries. VOSviewer software was used to represent the authorship networks, after normalizing the signatures and including those authors who had at least two deposited records.

3. Results
A total of 2,693 records were retrieved in the DCI containing 21,025 signatures and 2,679 authors were identified. Analysis of the repositories showed that the deposited data were mainly genetic in nature, with Gene Expression Omnibus standing out, followed by European Nucleotide Archive and Zenodo, with 2,414, 205 and 53 datasets, respectively (Figure 1). These results are corroborated by the Web of Science subject categories of the repositories where the datasets were deposited: Genetics & Heredity (2,623 datasets), Biochemistry & Molecular Biology (2,260 datasets) and Multidisciplinary Sciences (53 datasets) (Figure 2). Among the
most frequent Spanish authors in the practice of data deposition, Miguel Piris stood out with 306 records and links to 35 authors. On the other hand, Gonzalo Gómez López (183 records) had the highest number of links with other authors (n=113) (Figure 3). In turn, the main institutions involved in the data deposit were the Centro Nacional de Investigaciones Oncológicas (n=710), the Centro de Investigación Médica Aplicada (n=420) and the Centro de Investigación del Cáncer (n=196). Among the main collaborating countries with Spain in the practice of data sharing were the United States, the Netherlands and Australia.

4. Conclusions
The search equation designed in cancer generated in the last decade 2,693 research datasets with Spanish affiliation deposited in the Data Citation Index. Among the repositories used, those specialized in “omics” approaches predominate including genomics, proteomics, and metabolomics data. Compared to the deposited data, the same cancer strategy allowed 50,776 scientific publications to be obtained in the Science Citation Index Expanded. These results show that we are still far from the open science scenario demanded by Spanish, European and international public funding institutions, so there is a great need to make progress in understanding and overcoming the barriers that affect all those involved in the scientific process: researchers, publishers, and scientific organizations.

Figure 1. Classification of Repositories

![Classification of Repositories](image1)

Figure 2: Frequency of Web of Science Categories

![Frequency Web of Science Categories](image2)
5. Bibliographic references


Open science practices
Regarding the data obtained at the end of the development of this project, the preservation of the data generated with public financing will be guaranteed and can be reused. With this objective, the data will be deposited in the European repository Zenodo to make them available at the national and international level.
Furthermore, if this contribution is accepted for presentation at STI 2023, we fully agree with its open deposit through Zenodo or any other scientific platform for the same purpose.

Acknowledgments

Author contributions
All listed authors have made substantial contributions to the conception or design of the work; or the acquisition, analysis, interpretation of data and drafting of this work.
**Competing interests**
Authors declare no competing interests.

**Funding information**
This work benefited from assistance by the Ministerio de Igualdad de España (MUJER-PI-41-2-ID22), Ministerio de Economía y Competitividad del Gobierno de España (PID2019-108579RB-I00), Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital. Generalitat Valenciana (CIAICO/2021/205); Convocatoria 2015-Redes de Excelencia (CSO2015-71867-REDT); and Ministerio de Educación y Formación Profesional (MS21-020).