

BSC DATAVERSE

An Institutional Repository for Research Data and HPC

Paula Checchia (paula.checchia@bsc.es) and Aina Jené (aina.jene@bsc.es) - Barcelona Supercomputing Center (BSC-CNS)



Launched
April 2025



28
Collections



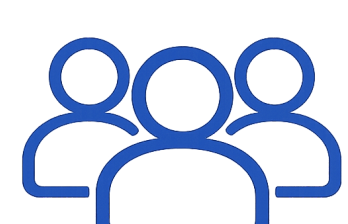
86
Datasets



6,327
Files

BSC Dataverse is the new BSC official **institutional research (meta)data repository**. It is designed to facilitate the **centralization, discovery, sharing, and reusability** of research data generated by BSC researchers, collaborators, and affiliated projects.

1 BSC CONTEXT BSC in numbers



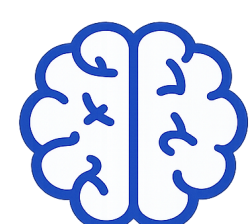
+1400
employees



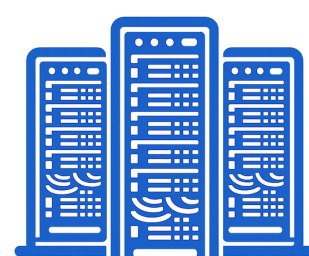
7
research
departments + laboratories



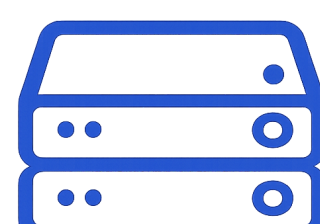
+70
research
groups



AI
Institute



MN5
supercomputing
resources



+650 PB
data stored
in house

2 THE CHALLENGE: OUR DATA IS DIFFERENT

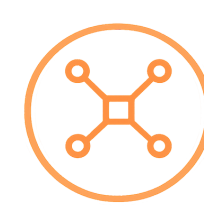
The problems



Datasets can be terabytes (not suitable for direct upload)



Data often lives in HPC storage or specialized external repos



Diverse data types
(= diverse metadata fields, standards, and tools)



Manual curation at scale is unsustainable

The solution

BSC Dataverse as a comprehensive institutional catalog.

- ✓ Flexible strategy adapts to each department
- ✓ Hosts files directly or links to external storage
- ✓ All BSC research findable in one place

3 DECENTRALIZED STEWARDSHIP MODEL

Each department/Laboratory has dedicated **Stewards** responsible for curation and data quality at the source.



4 FLEXIBLE STORAGE SYSTEM

PATH A - DIRECT HOSTING



Files stored directly in BSC Dataverse local storage

- ✓ Best for smaller datasets
- ✓ Self-contained & citable
- ✓ Direct download via Dataverse

PATH B - METADATA HUB



Metadata in Dataverse, files in external storage

- ✓ HPC/large datasets (TBs)
- ✓ Links to specialized repos
- ✓ Still FAIR - findable and citable



All BSC research is findable in one place

5 USE CASE: EARTH SCIENCES COLLECTION

Earth Sciences department

- 5** groups
- 1-2** data stewards per group
- 13** datasets published
- 2** sub-collections (permalink vs DOIs)

CLIMATE & AIR QUALITY DATA
Reanalysis datasets, ensembles of historical simulations, regional air quality datasets.

SOCIAL & POLICY DATA
Survey results, user requirements, socioeconomic data, text data.



Earth Sciences collection

Dataset examples

Path A example: CLIMA-GAP



Objective: Develop a RAG system to assess key characteristics in climate services projects and identify their contribution to success.



Webpage
(password: dataviz)



Dataset
(to be published)

- CONTENT OF THE DATASET**
Files required to reproduce the study
- Data**
 - Project list with links to PDFs
 - Prompts used
- Code**
 - PDF processing script + VDB creation
 - Inference script
- Outputs** (txt files with results by characteristic)

Path B example: HANAMI



Objective: Establish a standardized and validated methodology to test the replicability of climate models across different HPC platforms.



Website



Dataset
(published)

- CONTENT OF THE DATASET**
Links to the information required to reproduce the published paper
- Dataset description**
- 4 links to Zenodo**
- 1 script**
- 3 data repos** (simulations, NOAA & ERA5 data, etc. ~ 2TB per repository)



Check the BSC Dataverse using this QR code!
Or contact us at:
datamanagement@bsc.es



«La publicación es parte de la actuación CNS2023-145143, financiada por MCIN/AEI/10.13039/501100011033 y por la Unión Europea «Next Generation EU»/PRTR»