#CSIC

Use Cases from biodiversity and medicine and implications of data openness

A short reflection based on the analysis of research ecosystems

Prof. Jesús Marco de Lucas

(jesus.marco [at] csic.es) **CSIC, SPAIN**

Workshop of the OECD's Committee for Scientific and Technological Policy

Revision of Recommendation concerning access to research data from public funding:

use cases of enhanced access to software, algorithms and workflows

Special thanks to Lara Lloret, David Rodriguez, Fernando Aguilar, Alvaro López from IFCA

OECD Conference Center

October 15th 2019 @PARIS

Disclaimer: all statements presented are based in the personal experience as researcher and do not state any institutional position

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

CSIC

#CSIC

Understanding the research ecosystem

Who eats whom (what is the biomass? money? knowledge!)

SCIENCE IS IN THE BUSINESS CORE IN OUR CENTURY !

Science is no longer only the main interest of scientists…

GOVERNMENT AND PUBLIC ADMINISTRATION

PRIVATE AND PUBLIC COMPANIES

PUBLIC RESEARCH FUNDING & PERFORMING ORGANIZATIONS

-research (academic system)

-libraries and repositories

-technology transfer offices

-students

-citizens

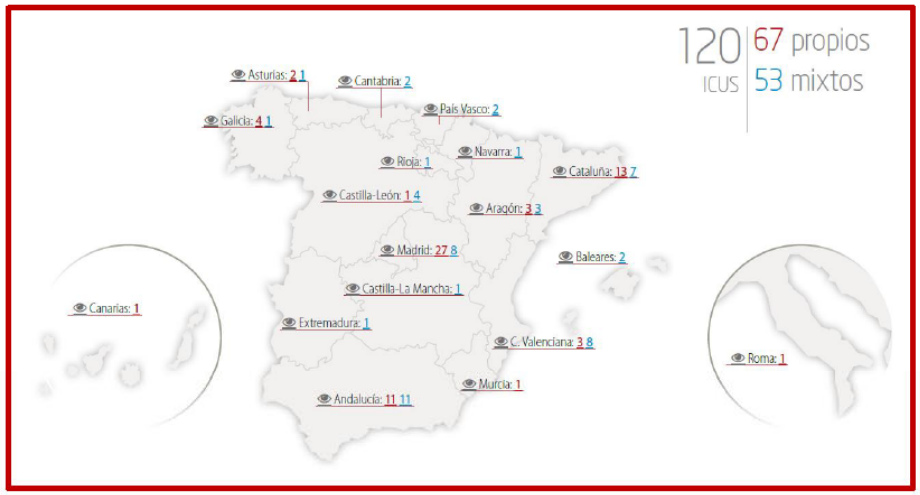
SOCIETY

*"Publishing journals as a **profit-maximizing business** is certainly as legitimate as it is for other **distributors of digital content based on intellectual property protections**. The research enterprise and its stakeholders are responsible for the future of scholarly communication"*
*From "**Open Science by design**" NASEM*

CSIC
Consejo Superior de Investigaciones Científicas

Consejo Superior de Investigaciones Científicas (CSIC)

#CSIC



INDICADORES ANUALES: >13.000 artículos / >1.600 contratos / >125 patentes solicitadas

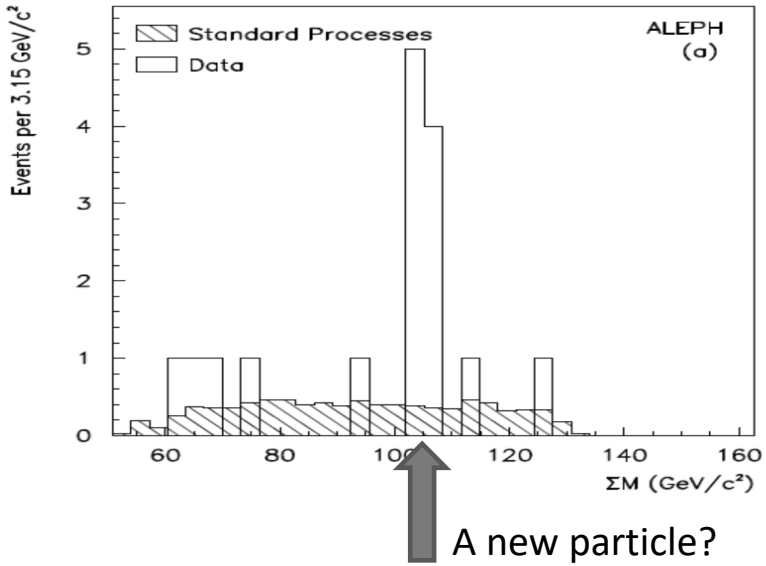
10.642 Empleados	3.644 Investigadores	1.263 Investigadores en formación	4.472 Personal de apoyo a la investigación	1.263 Gestión
5.220 mujeres	2.285 hombres	654 mujeres	2.445 mujeres	761 mujeres
5.422 hombres	1.360 mujeres	609 hombres	2.027 hombres	502 hombres

CSIC
Consejo Superior de Investigaciones Científicas

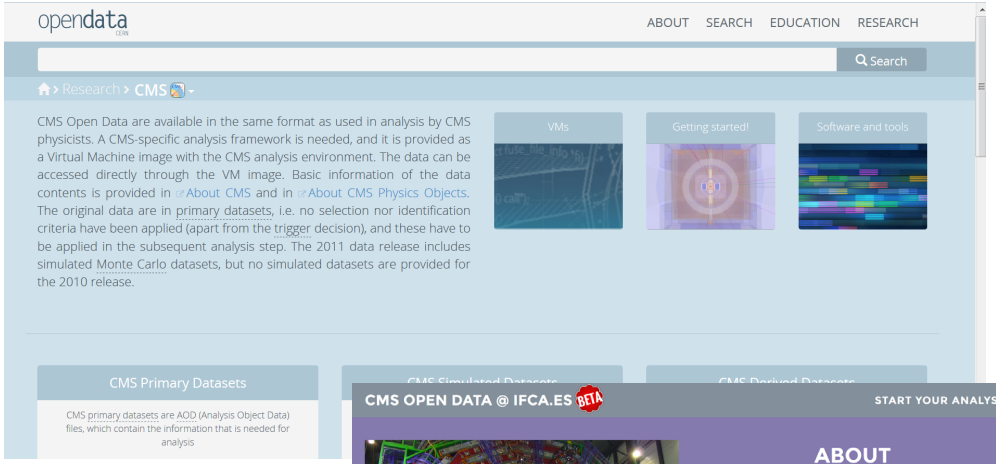
20 years towards open science in particle physics...

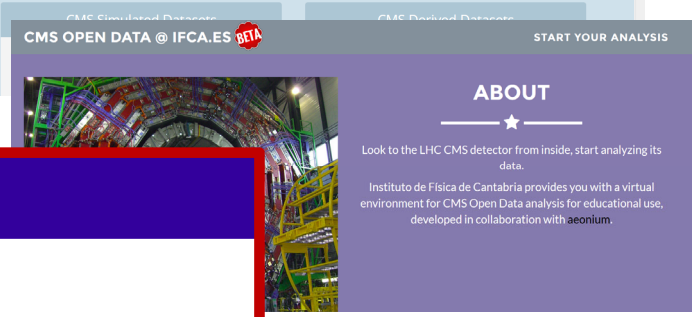
#CSIC *in one slide*

1995, LEP, CERN



2015, LHC, CERN





Many, many technical advances
Ecosystem: HEP community
Key factor: **REPUTATION**
Status: **PROGRESSING ADEQUATELY**

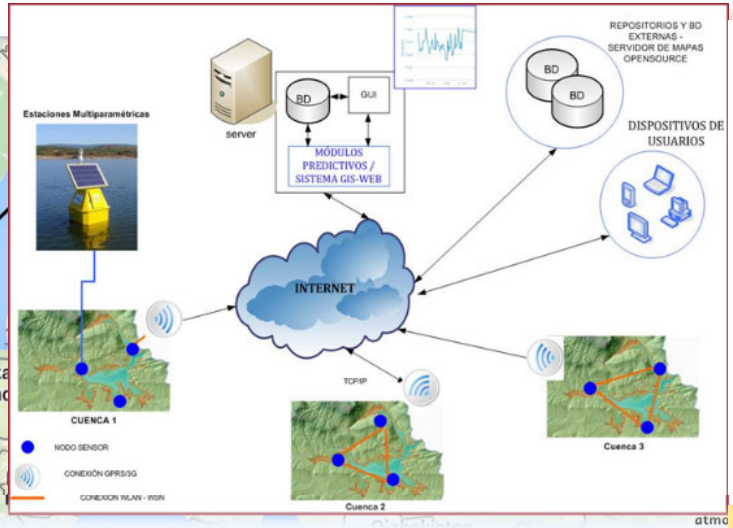
CSIC
Consejo Superior de Investigaciones Científicas

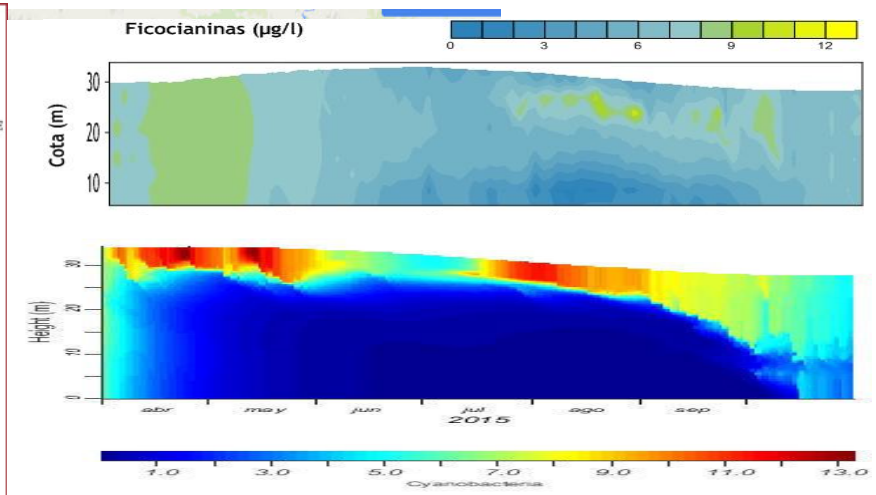
12 years project on Eutrophication coordinated by an SME

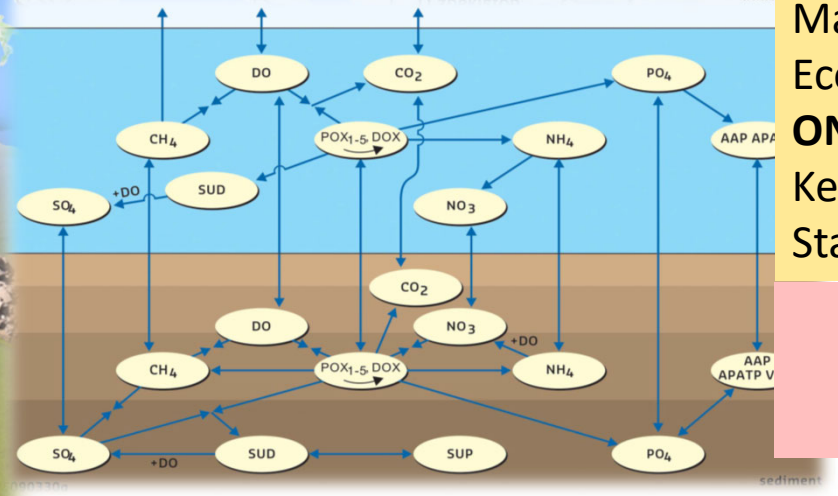
Ecohydros, SPAIN

#CSIC

Cuerda del Pozo Watershed (Soria, SPAIN)







Many, many technical advances

Ecosystem: **PRIVATE-PUBLIC COLLABORATION ON INTERDISCIPLINARY COMPLEX PROBLEM**

Key factor: **BUSINESS MODEL**

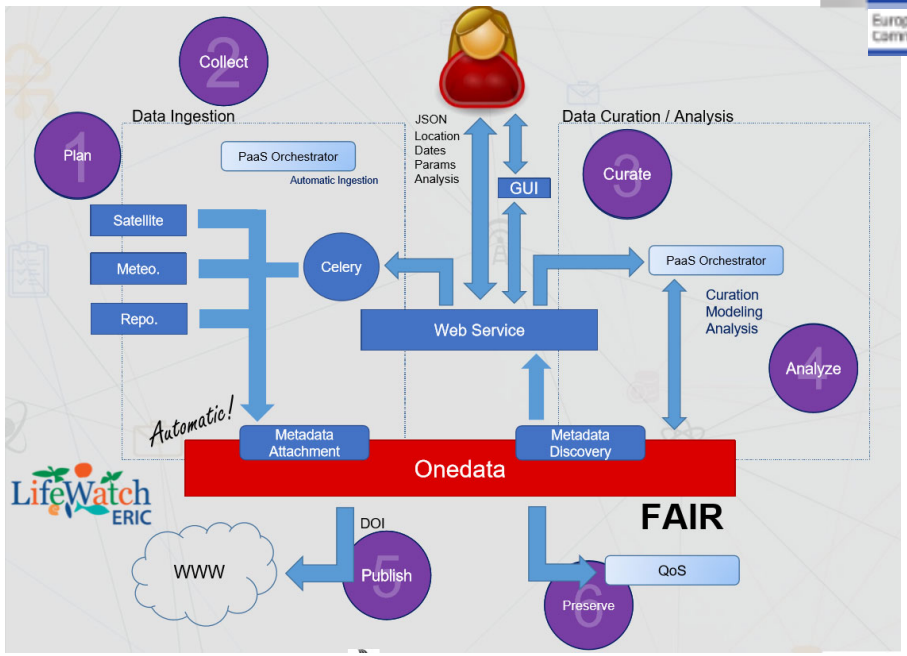
Status: **PROGRESSING ADEQUATELY?**

EXPLOIT SENTINEL DATA ON GOOGLE EARTH?

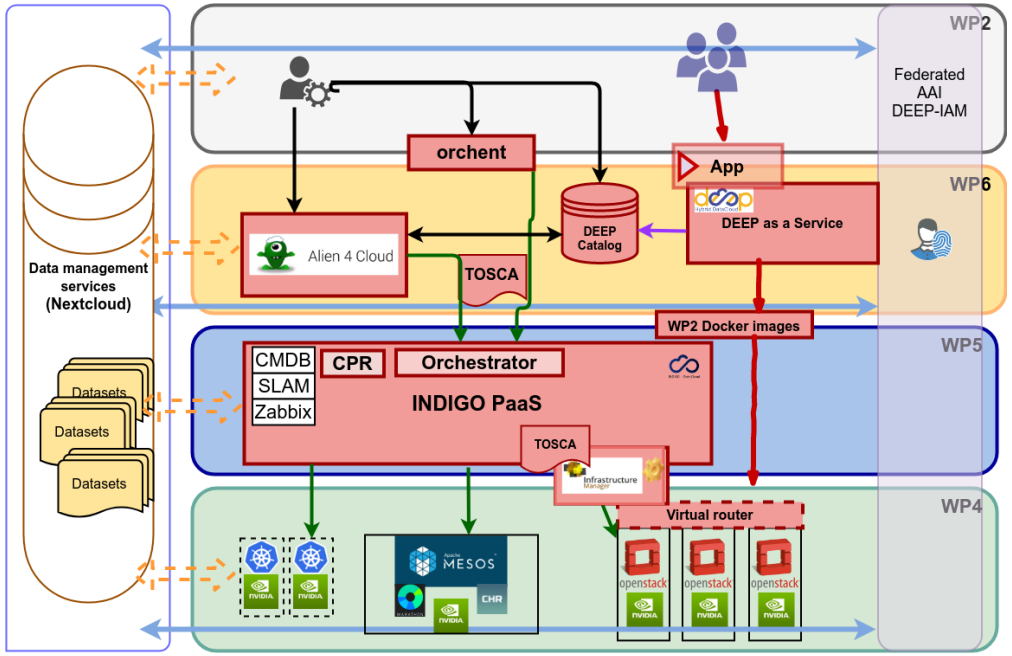
CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

Research Data Life Cycle and Cloud Platforms



F. Aguilar, XDC H2020 project



A. López, DEEP Hybrid DataCloud H2020 project

implemented following

A set of Common Software Quality Assurance Baseline Criteria for Research Projects

https://github.com/indigo-dc/sqa-baseline

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

Examples: understanding data and AI for research

Application to biology and medicine (examples)

IFCA Instituto de Física de Cantabria

- Successful projects developed on classification of plans and snails using a CNN.
- Some projects being started right now on CNN for medical imaging applications:
  - Mental diseases through nuclear magnetic resonances.
  - Fatty liver: Classification of liver biopsies
  - Brain hemorrhage through TAC images.

Exploiting same AI techniques:
From plants to plankton, to conus, to HEP collisions, to brain…

Ignacio Heredia

Arquitectura
ResNet50
Framework
Python con Lasagne/Theano
Training dataset
PlantNet *(6K especies | 250K imágenes)*

How to become the BEST expert in plant identification in your country in 10 minutes

- **Training dataset**
  Colección de imágenes de expertos (*68 especies | 1.5K imágenes*) que cubren tres regiones diferentes:
  - Región Panámica
  - Región de África del Sur
  - Atlántico Occidental y Mediterráneo
- **Resultados**
  Los resultados usando sets de imágenes de Google son prometedores

http://conus.deep.ifca.es/

Lara Lloret

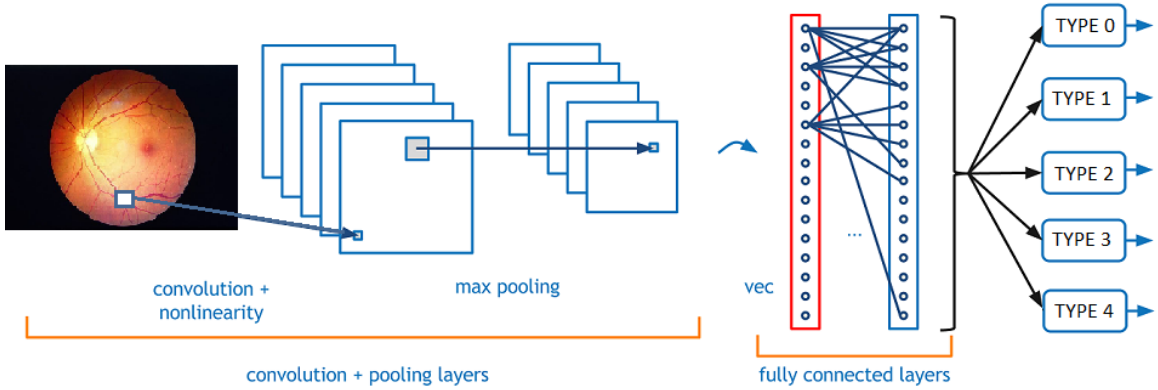
Do you think experts understand how they identify "**by heart**"?

CSIC Consejo Superior de Investigaciones Científicas

#CSIC

Research in Medicine: Joining Data and AI

- **Retinopathy** is a fast-growing cause of blindness over 400 million people at risk from diabetic retinopathy alone : successfully treated if it is detected early.
- Colour fundus retinal photography is analysed in order to document the presence of disorders and monitor their change over time.
- Specialized medical experts interpret such images and are able to detect the presence and stage of retinal eye disease such as diabetic retinopathy.
- **Deep Learning approach**: automated classification based on color fundus retinal photography images



Only an example *(but one of those few already accepted)* of **medical applications** where AI **should** help for good!

ECOSYSTEM:

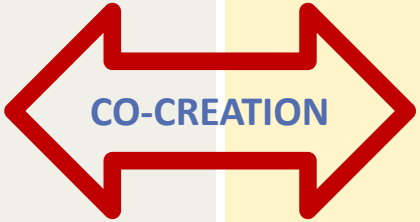
-biomed researchers /
-doctors
-patients
-biomed companies
-authorities
-hospital owners
-assurance companies
-families

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

A proposal: Joining Data and AI in a public safe environment

- **Donation of Medical/Personal Data**
  - Clinical History
  - Activity, Exercise, Food, Environment
- **Data Safe Haven** ***(+ HPC framework)***
- **Legal Sandbox**
  - Joint (Public-Private) Research proposals
  - Open Research Software Tools (on IA)
  - Joint (Public-Private) Exploitation



ECOSYSTEM:
-biomed researchers /
-doctors
-patients
-biomed companies
-authorities
-hospital owners
-assurance companies
-families

Many, many technical advances
Ecosystem: **PRIVATE-PUBLIC COLLABORATION ON INTERDISCIPLINARY COMPLEX PROBLEM**
Key factor: **PRIVACY**
Status: **HOW TO PROGRESS?**

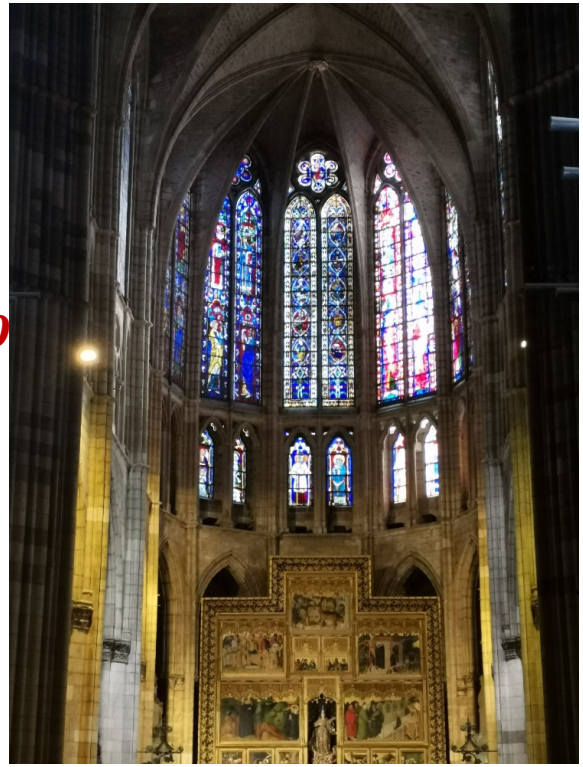
EXPLOIT MEDICAL DATA ON *"PUBLIC"* HPC CLOUD?
(IBM, Amazon, Google, Microsoft…)

CSIC
Consejo Superior de Investigaciones Científicas

What next from CSIC

#CSIC

- **Better understand each ECOSYSTEM** in Open Science, and our role
  - NOT ONLY MONEY BUT KNOWLEDGE (AND INFLUENCE) CONTROL
  - POSITION AT RELEVANT FORA WITH ARGUMENTS AND EXPERIENCE
  - DISCUSS IN DETAIL IP RIGHTS AND LICENSES (**like N-C on data**)
- **Reinforce support to Open Access/Science**
  - **CSIC Open Access Mandate (effective 1st April 2019)**
  - EXPLOIT EOSC ADVANCES ***(from EOSC-Hub, EOSC-Synergy, INDIGO DataCloud, XDC, DEEP Hybrid DataCloud, Cos4Cloud)***
  - **E-INFRASTRUCTURE** AND EVOLUTION OF P-P AGREEMENTS
  - **REINFORCE TRAINING AND DISSEMINATION, NETWORKING!**
- **Support new ideas exploiting Open Access/Data/Software**
  - **Data with privacy constraints but key for improving life quality**
  - Long term data series
  - Involve society (education and dissemination)
  - Keep in mind "AI FOR GOOD"



CSIC
Consejo Superior de Investigaciones Científicas

Answering Alan's question

#CSIC

- **Should we broaden the scope of the Recommendation to software, algorithms and workflows? If so, how?**
  - *Yes, software, algorithms and workflows need to be included, and this can be done relatively straightforwardly by adapting some of the provisions of the Recommendation so that they can **apply to both data and software***
  - *Yes, software, algorithms and workflows need to be included, but it needs much more work to build a consensus, and thus our decision could be to go forward on the current revision with data only, and work on software, algorithms and workflows in the 2021/22 biennium*
  - *No, building consensus around software, algorithms and workflows is not likely to materialise in the near future so we drop it.*
- **My answer:**
  - Yes, **BOTH data and software** (including algorithms and workflows) need to be included, but **BOTH** need more work to build a consensus. Answering to **"as open as possible"** implies the analysis of the research ecosystem and the corresponding exploitation (licensing) schemes.
  - **DIFFERENT SOLUTIONS ARE REQUIRED FOR DIFFERENT RESEARCH ECOSYSTEMS**
  - **A GENERAL RECOMMENDATION ON HOW TO AGREE ON SUCH SOLUTION WOULD HELP A LOT (USE CASES)**
  - **INDEPENDENTLY, QUALITY METHODS GUARANTEEING DATA AND SOFTWARE FAIRNESS SHOULD BE IMPLEMENTED (FOR THE BENEFIT OF THE RESEARCH ECOSYSTEM)**

CSIC
Consejo Superior de Investigaciones Científicas

#CSIC

Epilogue

« Je vis comme je peux dans ce pays malheureux, riche de son peuple et de sa jeunesse, provisoirement pauvre dans ses élites, lancé à la recherche d'un ordre nouveau et d'une renaissance à laquelle je crois.
Sans liberté vraie et sans un certain honneur, je ne puis vivre.»

« J'ai toujours pensé que si l'homme qui espérait dans la condition humaine était un fou, celui qui désespérait des évènements était un lâche. »

Albert Camus, 1958

CSIC
Consejo Superior de Investigaciones Científicas