Cancer Imaging Data FAIRification in EUCAIM

DAVID RODRÍGUEZ GONZÁLEZ
COMPUTACIÓN AVANZADA Y E-CIENCIA
INSTITUTO DE FÍSICA DE CANTABRIA (CSIC)

FERNANDO AGUILAR
CENTRAL ORGANISATION (VICYT)
SPANISH NATIONAL RESEARCH COUNCIL (CSIC)

IBERGRID 2023, Benasque

The EUCAIM Project

- **EU**ropean Federation for **CA**ncer **IM**ages: cornerstope of EU European Cancer Imaging Intiative part of the Europe's Beating Cancer Plan (EBCP)
  - 4 years Project started in January 2023
  - 76 partners
- Pan-European digital federated infrastructure of **FAIR** cancer-related, de-identified, real-world images.
  - Preserve data sovereignity
- Atlas of Cancer Images
  - Development and benchmarking of AI tolos towards preceision medicine

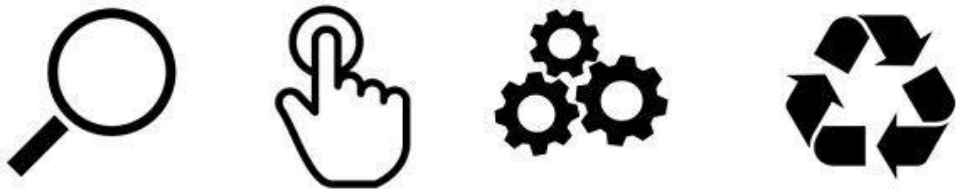
EUCAIM CANCER IMAGE EUROPE

The EUCAIM Project

- Builds on previous "Artificial Intelligence for Health Imaging" (AI4HI) projects
  - EuCanImage
  - Pro-Cancer-I
  - CHAIMELEON
  - PRIMAGE
  - INCISIVE
- Starts with 21 clinical sites from 12 countries
  - By 2026 at least 30 distributed data providers from 15 countries.

EUCAIM CANCER IMAGE EUROPE

Findable Accessible Interoperable Reusable



- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

F	F1	F1-01M	Metadata is identified by a persistent identifier	Recommended
	F1	F1-02M	Metadata is identified by a universally unique identifier	Recommended
	F1	F1-01D	Data is identified by a persistent identifier	Mandatory
	F1	F1-02D	Data is identified by a universally unique identifier	Mandatory
	F2	F2-01M	Sufficient metadata is provided to allow discovery, following domain/discipline-specific metadata standard	Recommended
	F2	F2-02M	Metadata is provided for the discovery-related elements defined by the RDA Metadata IG, as much as possible and relevant, if no domain/discipline-specific metadata standard is available	Recommended
	F3	F3-01M	Metadata includes the identifier for the data	Mandatory
	F4	F4-01M	Metadata or landing page is harvested by general search engine	Recommended
	F4	F4-02M	Metadata is harvested by or submitted to domain/discipline-specific portal	Recommended
	F4	F4-03M	Metadata is indexed in institutional repository	Recommended

RDA
RESEARCH DATA ALLIANCE

EUCAIM
CANCER IMAGE EUROPE

IFCA participation in FAIR tasks

- Complex Project
  - Different countries with different interpretations of GDPR
- FAIR related tasks (at least) in three WPs: 2, 4 and 5
  - Input from WP3 legal.
  - InterWP working group on FAIR metadata
  - Enable federated queries
    - Metadata catalogue based on AI4HI on
- We lead subtask 5.3.5 Data *FAIRification*
  - We also participate in T2.4 FAIR implementation support

T5.3.5 DATA Fairification (IFCA)

- FAIR compliance
- Define specific FAIR attributes for Cancer Imaging
  - Follow RDA recommendations
  - Dataset level metadata
    - Privacy concerns
- Evaluate tools and services for data “fairification”
  - Recommendations
- Leverage FAIR evaluator (EOSC-Synergy) for monitoring compliance

Federated Catalogue: AI4HI projects

- **EuCanImage**: MOLGENIS platform (https://doi.org/10.1093/bioinformatics/bty742), (https://www.molgenis.org/)
- **Pro-Cancer-I**: MOLGENIS platform.
- **CHAIMELEON**: Custom service called ¨Dataset Explorer¨, implemented at UPV. Datasets with ¨public¨ status are shared via ZENODO (including a DOI).
- **PRIMAGE**: There is not really a dataset catalogue.
- **INCISIVE**: There is no really a dataset catalogue.

Evaluation tools

- FAIR Eva (EOSC-Synergy.
- F-UJI
- FAIR Evaluator tool (ELIXIR)
- Other

Evaluation tools

- **FAIR EVA (EOSC-Synergy).**
- F-UJI
- FAIR Evaluator tool (ELIXIR)
- Other

Data FAIRification

- Domain-relevant metadata requirements
  - Machine actionable metadata components
- Go FAIR machine-actionable FAIR implementation profile (FIP)
  - https://www.go-fair.org/how-to-go-fair/fair-implementation-profile/
  - Identify identifiers for metadata
  - Identify identifiers for datasets
  - Metadata schemas
  - ….

Cancer Imaging specific

- Granularity
  - Federated catalogue
  - Trade-offs
- Existing ontologies
- Heterogeneity of sources
  - Find a minimum common denominator
  - Effort to make data FAIR
- Types of metadata the could/should be taken into account:
  - Clinical
  - Imaging (DICOM, modalities)
  - Patient Demographics? Reidentification risk

Clinical metadata

- EUCAIM Catalogue
  - Export data from other catalogues
  - Transformation step for encoding
- Body part examined
  - DICOM
  - Catalogues?
- Use coded information
  - ICD-10 (11)
  - SNOMED-CT
  - How much data already?
  - Existing tools

Imaging Metadata

- Modality
  - Acquisition details
  - E.g. MR sequences….
- DICOM
  - SOP Class UIDs
  - Machine readable
  - Need human-interface
  - Many more attributes present (Scanning sequence, options,….) which are necessary?

FAIR EVA – Evaluator, Validator & Advisor

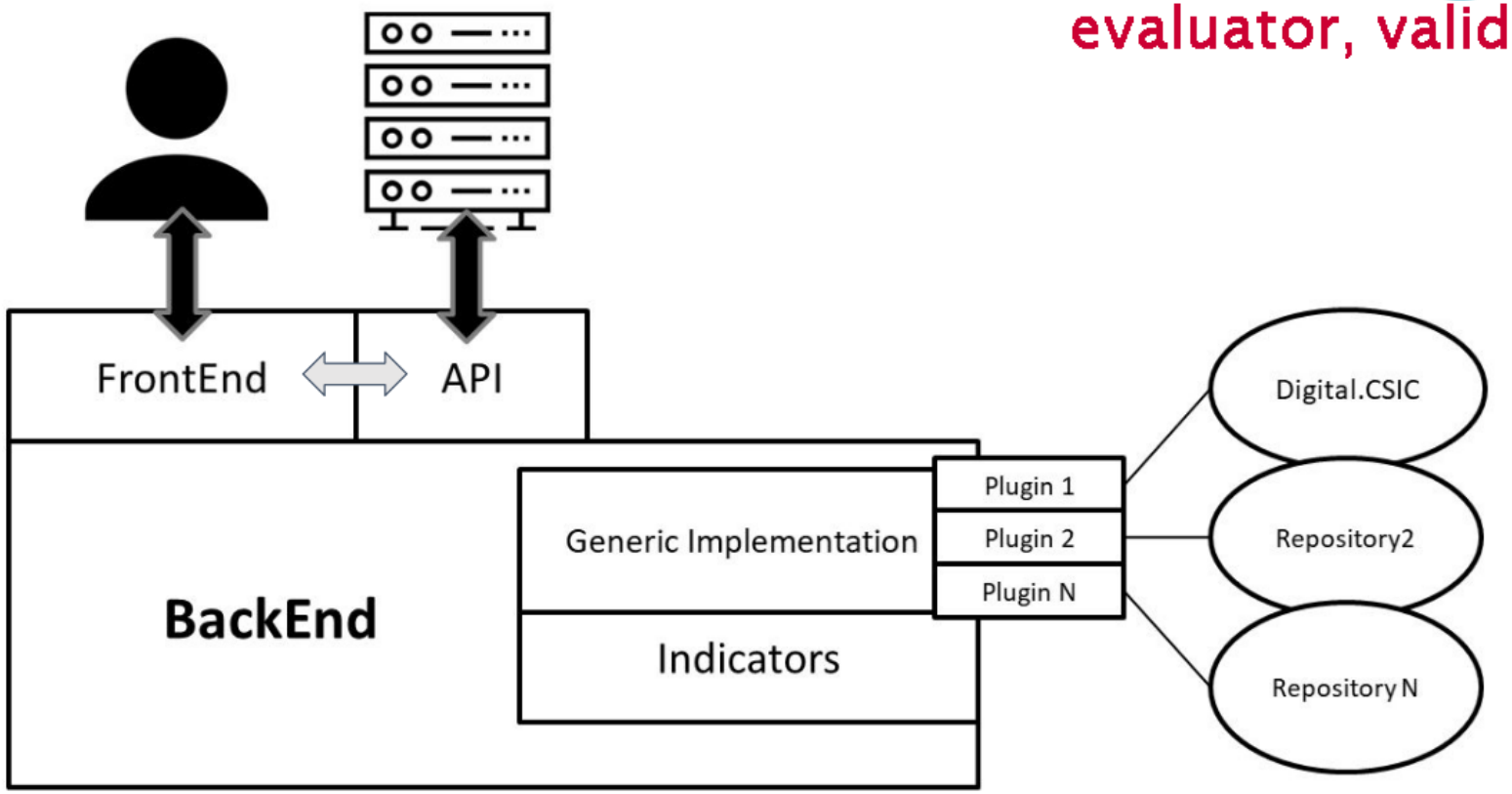
- EOSC-Synergy Project's product
- FAIR indicators technical implementation
  - Starting from RDA
- Modular, Scalable, Flexible
  - Generic Implementation
  - Plugins
- Not only evaluate, but also validate and advise
- Python API + web interface
- Stand-alone Docker

FAIR Assessment tools - FAIR EVA

- FAIR EVA functionality
- Comply with **FAIR Data principles**:
    1. **Data**: use a proper format
    2. **Metadata**: community standard. Machine-actionable (JSON, XML, RDF...)
    3. **PIDs**: Persistent Identifier (e.g. DOI). Provided by an accepted authority.
    4. **Repository/Data service**: indexed and machine-actionable.
- Integration: Different types of repositories/data portals/repository softwares

FAIR Eva Architecture

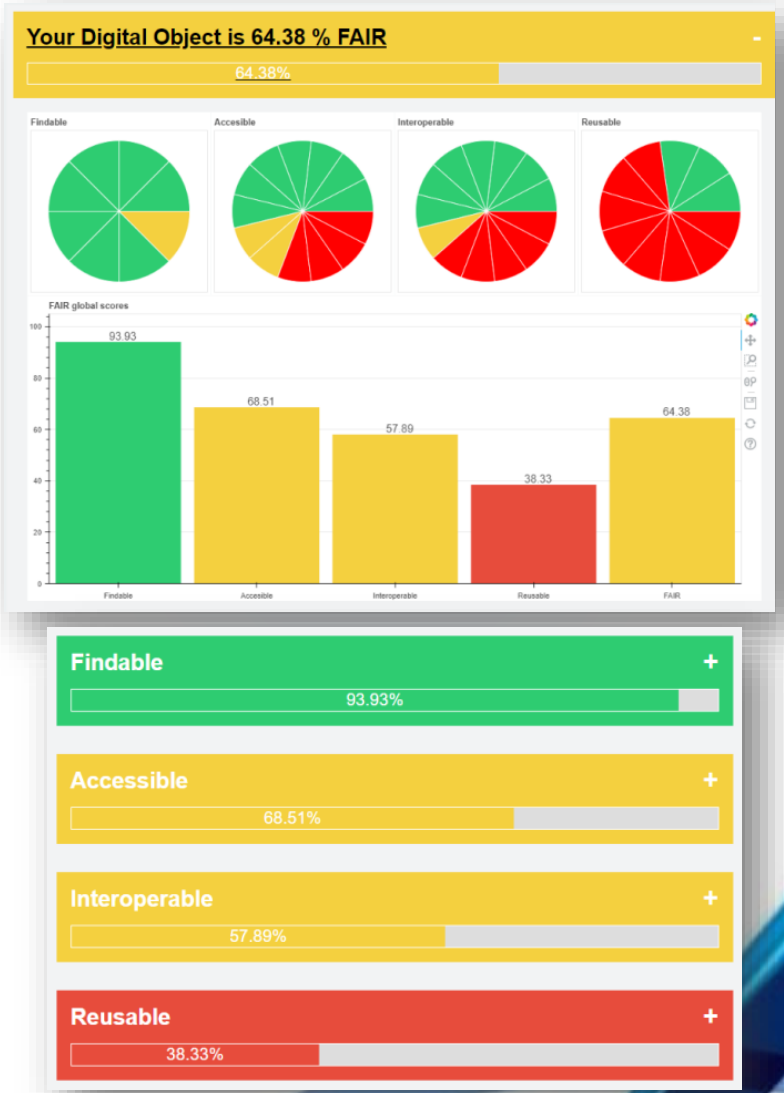
FAIReva
evaluator, validator & advisor



FAIReva

evaluator, validator & advisor

- Open source:
  - git clone https://github.com/EOSC-synergy/FAIR_eva.git
  - webinar: https://www.youtube.com/watch?v=YhSPUYbqIdo
- Developed in Python
- Stand-alone - Docker deployment
- DIGITAL.CSIC Beta: fair.csic.es
- Plugins on development - GBIF, DT-GEO (EPOS ESFRI)



EUCAIM CANCER IMAGE EUROPE

Next steps

- First versión of EUCAIM specific metadata FAIR attributes being defined
  - Also the levels of compliance required for federated repositories
- Once agreed we need to implement check mechanism
- Develop a FAIR EVA plugin
  - Will take information from Molgenis
  - Test and feedback
  - Deploy in the central node
- Help other federated resources to adopt it, if they want
- Further iterations

Summary

- FAIR data is a mandate for the EUCAIM Project
- However the sensitivity of the data and the divergences in the interpretation of GDPR in different countries makes it difficult
- Complex project
  - Diversity of data sources also a challenge
- FAIR EVA chosen as testing tool for evaluating the level of compliance
- Plugin development necessary to add new checks specific to EUCAIM – Cancer Imaging

Questions

Thanks!

David Rodríguez González

drodrig at ifca.unican.es