National Czech Programme | eosc | MUNI ICS | BioMedAI TWINNING | RationAI | BBMRI-ERIC The European research infrastructure for biobanking and biomolecular resources in health and life sciences

From Samples to Algorithms: Building Infrastructure for Biomedical AI

Assoc. Prof. Petr Holub, Ph.D.

Chief Information Officer (CIO), BBMRI-ERIC

Head of BioMedAI, Institute of Computer Science, Masaryk University

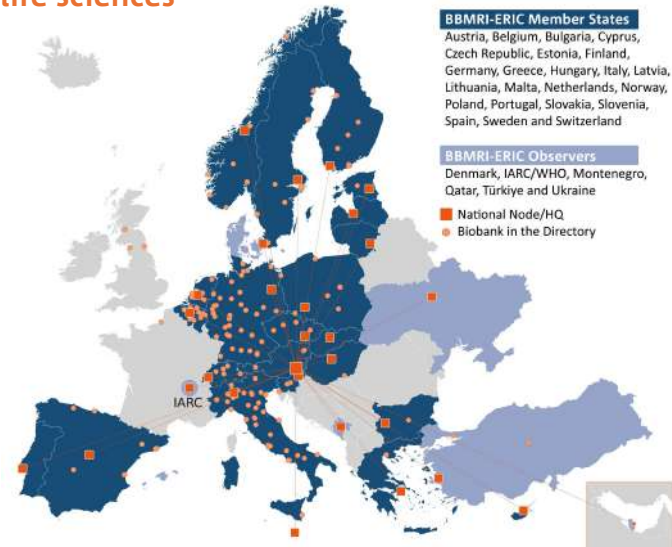
BBMRI IS A DISTRIBUTED ERIC

Operating in health and life sciences

BBMRI-ERIC®
The European research infrastructure for biobanking and biomolecular resources in health and life sciences

THE BBMRI COMMUNITY

~ 500 biobanks at clinics and universities across 32 countries
28 National Nodes + IARC/WHO
3 Biotechnology Expert Centres
>30 ELSI experts
>130 Information Technology experts
>300 Quality Management experts
1 Headquarter



SERVING SOCIETY, ACADEMIA AND INDUSTRY ALSO AS DATA INFRASTRUCTURE

THE LOCAL BIOBANK ECOSYSTEM

HEALTHCARE INTEGRATED BIOBANKS = **DATA INTEGRATION CENTRES**

CLINICAL CARE



TRANSLATIONAL RESEARCH



- Patient recruitment & CONSENT
- Sample & Data COLLECTION
- Sample & Data PROCESSING
- Sample & Data STORAGE
- Biobank Information Technology (IT) System
- Sample & Data SEARCH, REQUESTS, & TRANSFER

QM-Certified & internationally connected

In support of

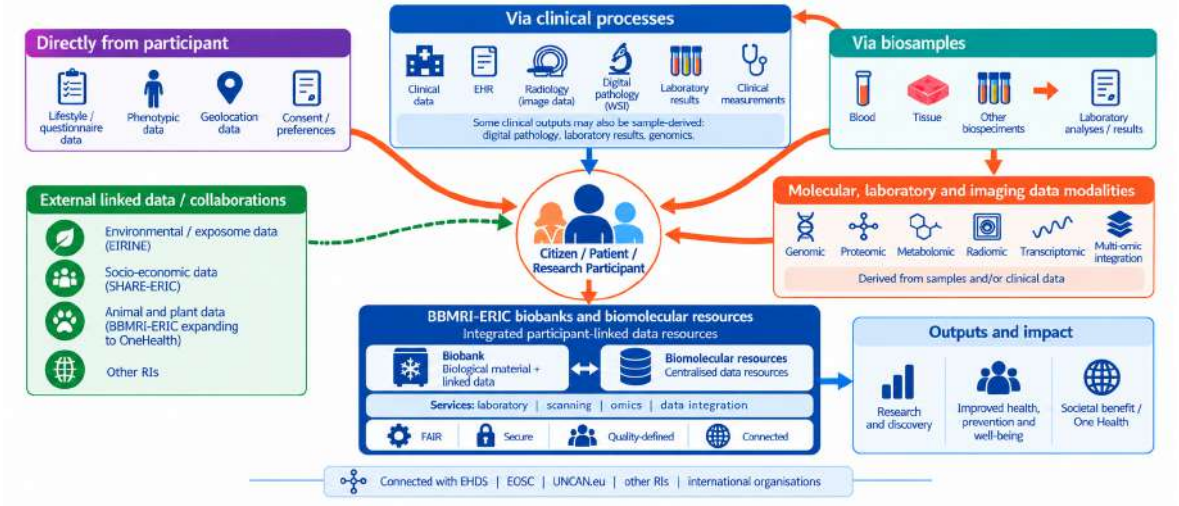
- Patient Empowerment
- Citizen Engagement
- Clinical Trials
- Molecular Tumour Boards

Enabling access to

- Sample, clinical, follow-up, and research DATA
- Biomolecular Resources (e.g., OMICs technology)
- Imaging
- Bioinformatics & AI tools

National Czech Programme | eosc | MUNI ICS | BioMedAI TWINNING | RationAI | BBMRI-ERIC® The European research infrastructure for biobanking and biomolecular resources in health and life sciences

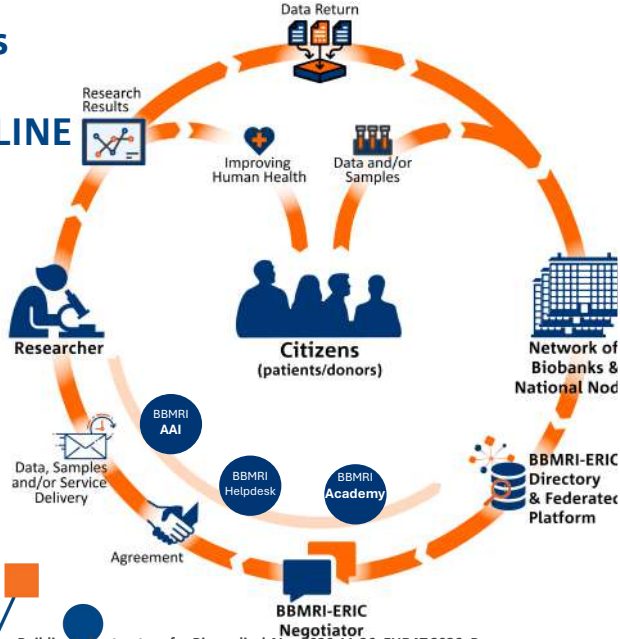
Data Modalities in BBMRI-ERIC



BBMRI-ERIC's FEDERATED ACCESS PIPELINE

BBMRI-ERIC Directory is largest biobank catalogue worldwide:
- ✓ 473 biobanks with
- ✓ 9.9 Mio. citizens across
- ✓ 33 countries.

BBMRI-ERIC Federated Platform features:
- ✓ 450,000 citizens with
- ✓ 155,000 genomes across
- ✓ 10 EU countries today.



BBMRI-ERIC®
The European research infrastructure for biobanking and biomolecular resources in health and life sciences

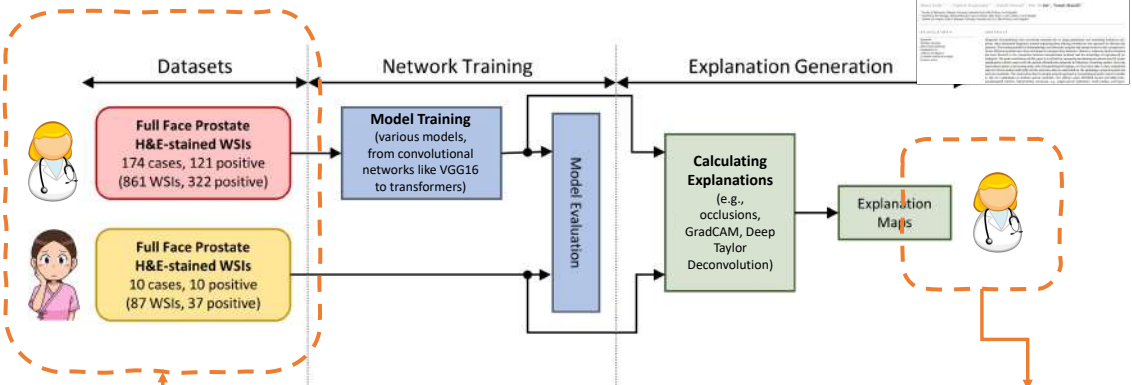
Selected KPIs 2024 (2023)

- **10,444 (8,924)** users of the Directory
- **9,134 (6,392)** successful access requests
- **3,204 (2,566)** publications involving BBMRI
- **42,992 (24,807)** participants in outreach events

BBMRI-ERIC & AI

- **BBMRI-ERIC as a data provider for clinically-relevant AI**
  - sample-derived data (imaging, omics, …) and associated data (clinical, phenotype, exposure, …) for **model training**
  - **environment to validate AI models on unseen data – single-centric or multi-centric**
- **BBMRI-ERIC as an expertise provider for clinically-relevant AI**
  - identification of clinical experts
- **BBMRI-ERIC as a host of AI groups – to improve services of the RI, e.g.,**
  - **quality checking**
  - **providing descriptors** for data sets **to facilitate discovery**
  - **generate synthetic data** to speed up or proxy access

How is BBMRI-ERIC used for XAI?



MCVAL EOSC Federation Use Case

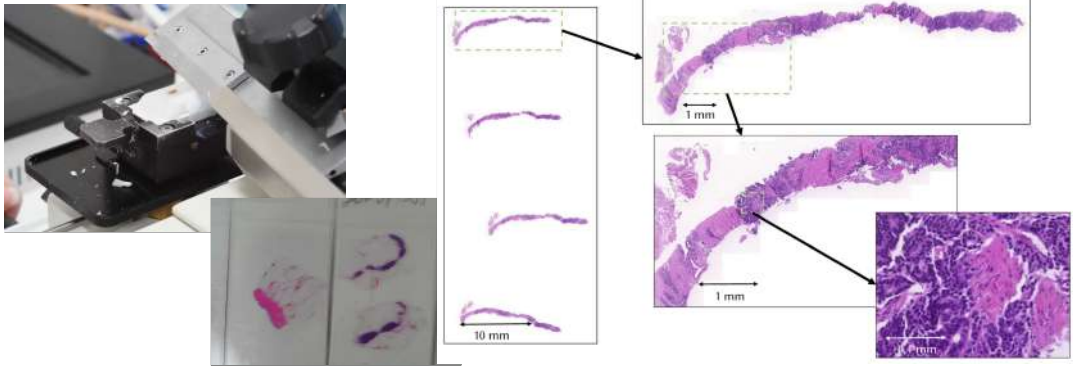
- **Evolution of pathology**: from examining dead bodies to **critical cancer diagnostics**
- Projected digital pathology growth: $1.2B (2024) to $2.6B (2032) (Delvens)



MCVAL EOSC Federation Use Case

Example of prostate slide

— Statistically: about 1 in 8 men will be diagnosed with prostate cancer during their lifetime (risk varies with age, race/ethnicity, other factors…).



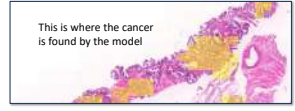
National Czech Programme | eosc | MUNI ICS | BioMedAI TWINNING | RationAI | BBMRI-ERIC® The European research infrastructure for biobanking and biomolecular resources in health and life sciences

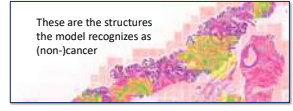
MCVAL EOSC Federation Use Case

- **Validation of prostate cancer detection model developed and used by Masaryk Memorial Cancer Center (BBMRI.cz host) and Masaryk University, on multi-centric European data**
  - in routine clinical use at MMCI after internal validation
  - significant exploration of explainability of the AI model
  - initially piloted with BBMRI.at/MUG data
  - **demonstrated in EOSC Symposium 2025**
- **Preliminary results**
  - 90% sensitivity, 70% specificity on MUG data compared to 100% sensitivity and 94% specificity at MMCI
  - all positive patients correctly identified on MUG data



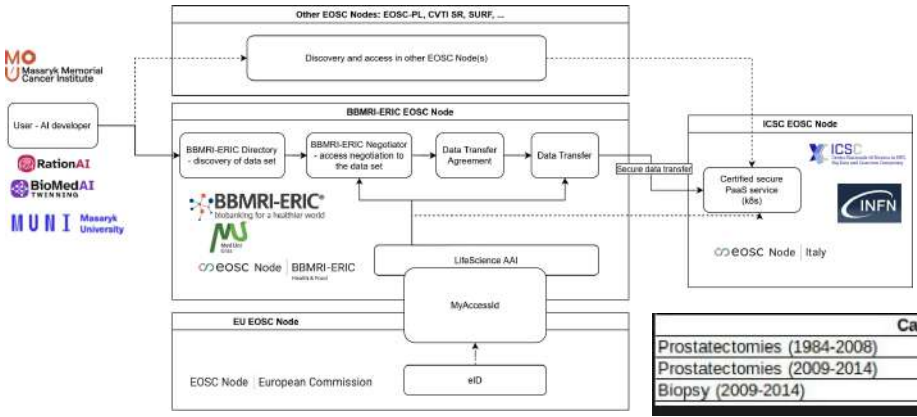






MCVAL Demonstrator in Practice

Roles – 36 experts involved!
- AI researcher
- Secure compute node
- Sensitive data provider
- AI validation service provider



Participants
- MMCI/MU
- ICSC/INFN
- MUG
- BBMRI-ERIC

Additional EOSC Federation Sensitive Data WG Participants
- EOSC-SK
- EOSC-PL
- EOSC-NL (SURF)
- EOSC-FI

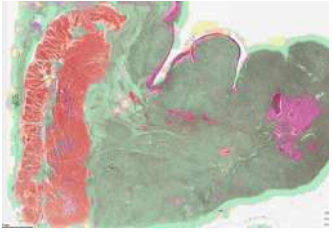
	Cases	Slides
Prostatectomies (1984-2008)	5525	104335
Prostatectomies (2009-2014)	794	28.786
Biopsy (2009-2014)	331	11.193

National Czech Programme | eosc | MUNI ICS | BioMedAI TWINNING | RationAI | BBMRI-ERIC® The European research infrastructure for biobanking and biomolecular resources in health and life sciences

Improving Discovery & Access

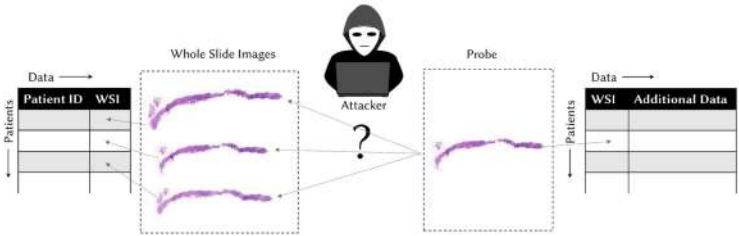


- **Tissue classification**
  - improving **findability in large archives of WSIs**
    - only Medical University Graz and Masaryk Memorial Cancer Institute have cca 3 PB (>500k) WSIs
    - archives with 50M+ histopathological slides could be scanned
  - **rough classification of tissue types and organs** on otherwise undocumented slides
  - **generating metadata descriptors** to enable findability
  - steer synthetic image generation
  - `https://doi.org/10.1016/j.artmed.2026.103368`



Data Anonymization and Synthesis

— **Domain-specific approaches – example of digital pathology data:**



— **Developed metohdology how to generate anonymized WSI data**
  — with low residual risk so that it can be shared as non-personal data

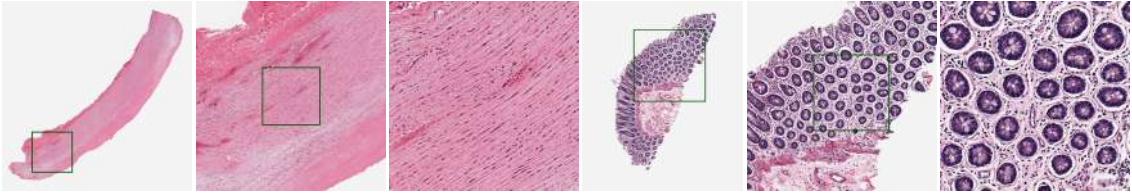
— **Guidelines to handle WSI as pseudonymized data**
  https://www.nature.com/articles/s41467-023-37991-y



Data Anonymization and Synthesis



- **Ability to fundamentally accelerate/pipline access if synthetic data is available**
  - generating temporary proxy data on demand to match real data
  - preserving privacy while synthesizing data
- **Recent work on gigapixel WSI data synthesis by Dr. Müller Group (MUG) – lead by Robert Harb**
  - multi-scale diffusion-based generative model



https://doi.org/10.48550/arXiv.2311.08199

Publishing AI Models in EOSC

BBMRI-ERIC will build on results of EOSC-CZ/Open Science II to support the following two workflows:

- **Publish and discover models across EOSC catalogues**
  - Give models persistent identifiers and link them to datasets and scripts.
  - Use Dublin Core and CCMM for descriptions and DCAT-AP for catalogue exchange.
- **Access and run models on local or remote data**
  - Run models on researchers' own data or on authorised data from another repository.
  - Enforce access conditions for both sensitive models and sensitive data.

Metadata and Licensing for Model Reuse

- **Is the model suitable for the task?**
  - Describe its task, architecture, evaluation results, intended use and limitations.
- **How was this model version produced?**
  - Link the base model, training data and scripts using PROV-O / ISO 23494.
  - Use Croissant to describe the structure of the training data.
- **Which components may be reused, and on what terms?**
  - Record machine-readable terms for reusing code, data, documentation and weights.
  - Intellectual property protection for model weights remains uncertain.

National Czech Programme

eosc

MUNI ICS

BioMedAI TWINNING

RationAI

BBMRI-ERIC®
The European research infrastructure for biobanking and biomolecular resources in health and life sciences

Thank you!

hopet@ics.muni.cz
petr.holub@bbmri-eric.eu

Co-Funded by the European Union

Ministry of Education, Youth and Sports of the Czech Republic

CC BY

Open Science II
Reg. n.: CZ.02.01.01/00/24_030/0015041