EUHUBS4DATA
EUROPEAN FEDERATION OF DATA DRIVEN INNOVATION HUBS

Experiments in action: the perspective of DIH supporting the experiments in practice

Marcin Plociennik

Poznan Supercomputing and Networking Center

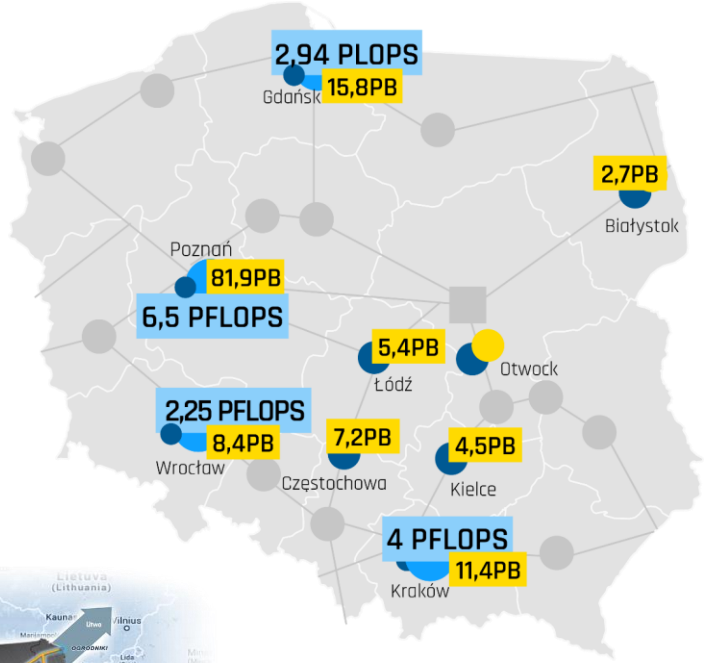
The EUHUBS4DATA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951771

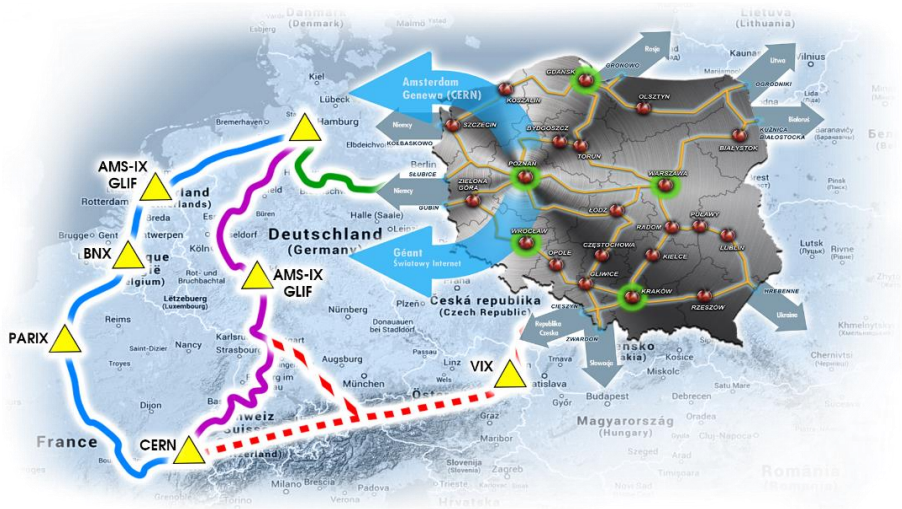
BIG DATA VALUE PUBLIC-PRIVATE PARTNERSHIP

This project is part of BDV PPP

PSNC – data driven infrastructure

- Leading **operator of Polish e-Infrastructure**
  - National Research and Education Network – PIONIER
  - Research Metropolitan Area Network - POZMAN
  - HPC Center: under upgrade
  - 2 Data Centers: over 80 PB under upgrade -> NDS





PSNC

EUH4D

PSNC – data driven infrastructure

National Data Storage infrastructure:

Planned:

250 PB tape (long-term storage/archive)

250 PB disk (on-line storage, data lake)

10 PB SSD/NVMe (access acceleration)

Target: **1 EB of storage capacity**

Partners: 5 HPC & 4 MAN sites:

HPC: PSNC, Cyfronet, TASK, WCSS, NCBJ

MAN: Białystok, Częstochowa, Łódź, Kielce

A complete environment for data-driven experimentation:

Data preparation & preservation:

- Data acquisition, cleaning, enrichment
- Discovery & search
- Sharing & publishing
- Data re-use, discovery and exploitation
- Data protection / preservation

Data processing & analytics:

- High performance computing
- Data Analysis, Data Science…
- Machine Learning, AI

PSNC as hub in EUhubs4Data

Services

SERVICE NAME	CATEGORY	DESCRIPTION	KEYWORDS
GEOSPATIAL TOOLKIT	Big Data Platforms (PaaS)	PSNC geospatial toolkit brings together a set of tools provided b...	GeoTools, GeoSpatial, map creation, geo-data
LINKED DATA PIPELINES SERVICE	Product Development	The main goal of these pipeline is to define and deploy (semi-) a...	data integration, linked (open) data, Ontologies, pipelines
ROHUB	Data Management	ROHub is a holistic solution for the storage, lifecycle managemen...	fair objects, research management, objects
SEMANTIC ANNOTATION	Product Development	The semantic annotation service provides a simple REST API design...	agriculture, Entity recognition, linke text mining
SPARK	Big Data Platforms (PaaS)	Apache Spark is general purpose, high performance and horizontall...	spark,machine learning
SYMBIOTE	Data Management	Symbiote is an ecosystem for connecting, exchanging information a...	IoT interoperability, sensor data

Mission

HPC4 Poland

Create and provide advanced high performance computing tools, addressing the real-life demand of Polish manufacturing companies

Datasets

OPEN LAND USE - OLU	Agriculture, Fisheries, Forestry and Food, Environ..	Creative Commons CCZero (CC0-1.0)	Czech Republic, Poland, Spain	Open Land Use Map is a composite map that is intended to create d...
OPEN TRANSPORT MAP - OTM	Agriculture, Fisheries, Forestry and Food, Transpo..	Creative Commons CCZero (CC0-1.0)	Czech Republic, Poland, Spain	The Open Transport Map displays a road network which – is suitabl...
POLISH LAND PARCEL INFORMATION SYSTEM (LPIS) DATASET	Agriculture, Fisheries, Forestry and Food, Environ..	Creative Commons CCZero (CC0-1.0)	Poland	Polish Land Parcel Information System (LPIS) data (land parcel an...
SMART POINTS OF INTEREST DATASET	Agriculture, Fisheries, Forestry and Food, Environ..	Open Data Commons Open Database License (ODbL-1.0)	EU-wide	The Smart Points of Interest dataset (SPOI) is the seamless and o...
URBAN ATLAS (LAND USE AND LAND COVER DATA FOR LARGE URBAN ZONES) DATASET	Agriculture, Fisheries, Forestry and Food, Environ..	Creative Commons CCZero (CC0-1.0)	Czech Republic, Poland, Spain	This dataset contains agriculture related lands (hilucs_code<200)...

PSNC

EUH4D

Open call experiments supported by PSNC

LONG RANGE PLANNER

Increase up to 160% the data acquisition capabilities of long-range drones in power line inspection thanks to developing an automatic mission planning

COMPANY	COUNTRY
FuVeX	Spain

AI4SAM

This experimentation will focus not only on the optimization problem of maximizing the reachability to the mobility means while minimizing the installation costs, but also on exploring incentivization models for fostering citizens engagement to soft ...

COMPANY	COUNTRY
KENTYOU	France

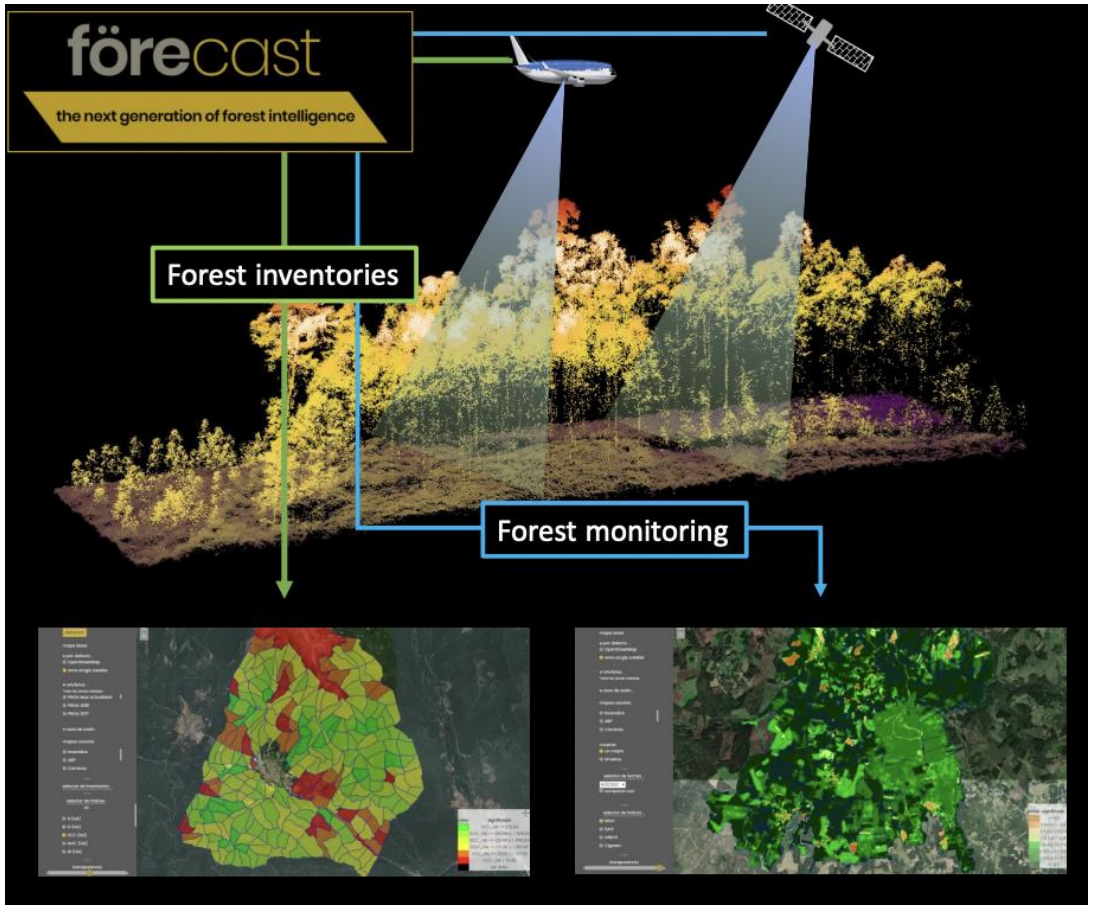
FÖRECAST 2.0

Förecast is a forest intelligence solution, designed to offer accurate and updated forest inventories in a multi-purpose platform. förecast provides real-time Earth Observation (EO) geo-information products by leveraging and integrating data from ...

COMPANY	COUNTRY
FORA FOREST TECHNOLOGIES, SLL	Spain

Success story - Förecast experiment

- **Förecast** provides real-time Earth Observation (EO) geo-information products by leveraging and integrating data from satellite, airborne LiDAR, aerial orthoimages and forest information to create **advanced algorithms for** high accuracy, reliability, up-to-date and high resolution to forest assets across different **ecosystem services**.
- **New** the design and operative **architecture: BigData Infrastructure**
- **New methodology to get non-parametric models** based on field truth and auto labelled.: **Geospatial scalability**



förecast
the next generation of forest intelligence

https://forecast.fora.es/

föra
forest technologies

info@fora.es
www.fora.es

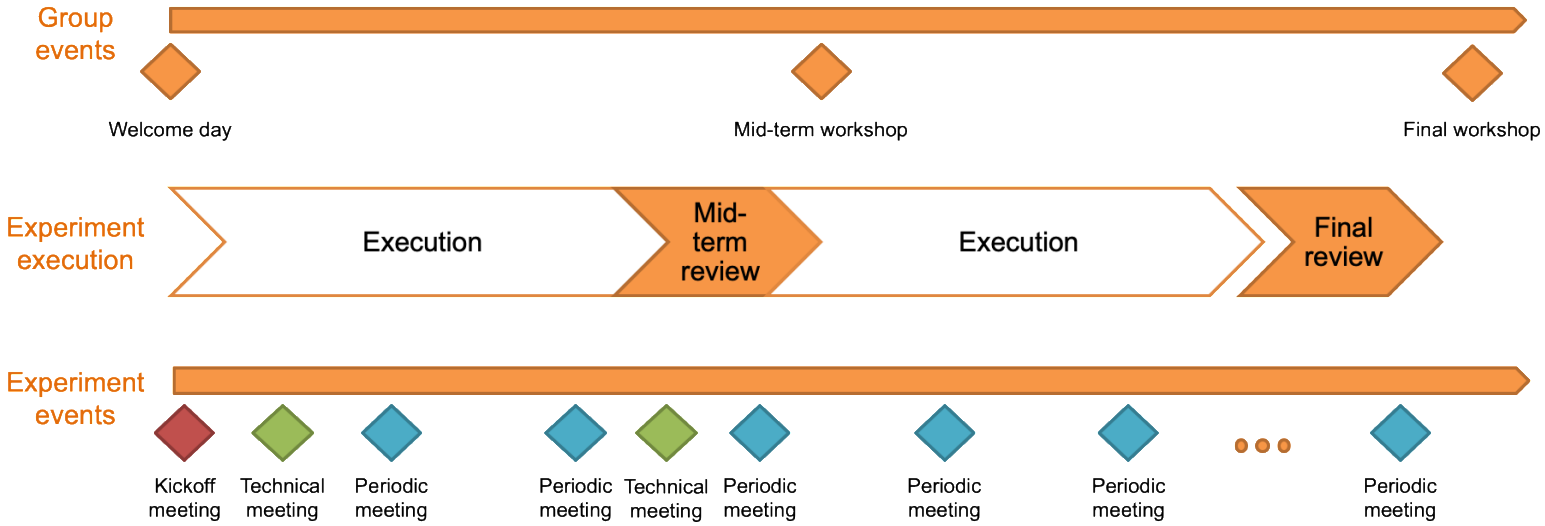
Forest technology and innovation company
Assessment of forest resources (wood, biomass, carbon, etc.) using remote sensing (LiDAR, satellite, etc.) and AI

DIH's Services supporting Förecast experiment

Experiment	Forecast 2.0	
SME	Fora Forest	
Coach	PSNC	
Service 1	DIH ARAGON	Consultancy service for business development and ecosystem building needed for optimal commercial roadmap of the service.
Service 2	EUT	Advisory role
Service 3	PSNC/HPC 4Poland	Geospatial Toolkit
Service 4	PSNC/HPC 4Poland	Data processing infrastructure

- **PSNC services used:**
  - **Geospatial toolkit**
    - Hadoop platform used for Geospatial Toolkit: 384 CPUs, 768 GB RAM, 21TB storage
  - **Cloud**: 11 nodes x 8cores x16GB RAM, 19.5 TB storage

Experimients lifecycle



- Ethical assessment and monitoring of the experiments (Ethics monitoring group in the project)

- Each experiments must provide a Data Management Plan (DMP)
  - **template** was provided **to get the information** related to the used and produced datasets

DATASET 1

Data set reference and name:	
Data set description:	*Description of the data that will be generated or collected*
Data utility:	*To whom the data could be useful and the reason why it is worth generating, preserving and/or sharing*
Type:	*Collected / Generated*
Nature:	*text, numbers, image, video, etc.*
Scale:	*the expected size of the dataset in MB/GB*
Origin:	*where does the data in the dataset come from, from which sources it has been collected*
Standards and metadata:	*Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created*
Data sharing policies:	*Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).*
Archiving and preservation:	*Identification of the repository where data will be stored, if already existing and identified. Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.*

Observations after the experiments round

Short term

- During experiments we had opportunity to strengthen not only collaboration with participating SME's but also between participating DIHs (DIHs ↔ DIHs) on the organizational and technical level
- Immediate feedback received from SME's: agile approach to new/changed requirements, improving/adjusting service offerings to the real scenarios

Long term

- Experiments are useful in **validating also the quality of services, datasets and support** offered by the Hubs and so, by the federation
- There are **high expectations** and demand in terms of the accuracy and **quality of the datasets** from SME's perspective as those datasets are having fundamental impact on the quality of the final SME's
- We had opportunity to attract some of the users assets into the federation, however it require lot of effort to create awareness and to show benefits for companies in sharing a data

PSNC perspective on the EUhubs4Data federation

Short term

- Common service and data catalogue
- Visibility and networking
- Wide offer of the trainings in federation
- Building blocks supporting data spaces

Long term

- Strong relations with ot*her DIH's*
- Access to wide range of services for the local ecosystem SME's
- Interoperability
- Trust building

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951771

Thanks a lot!

marcinp@man.poznan.pl

www.euhubs4data.eu