EUDAT

Towards a pan-European Collaborative Data Infrastructure

Damien Lecarpentier
CSC-IT Center for Science, Finland
CESSDA workshop
Tampere, 5 October 2012

SEVENTH FRAMEWORK PROGRAMME

EUDAT

EUDAT

Towards a pan-European Collaborative Data Infrastructure

Damien Lecarpentier
CSC-IT Center for Science, Finland
CESSDA workshop
Tampere, 5 October 2012

SEVENTH FRAMEWORK PROGRAMME

EUDAT

European Data

EUDAT

- Start date: 1st October 2011
- Duration: 36 Months
- Budget: 16.3 M€ (9.3M€ EC)
- EC Call: INFRA-2011-1.2.2
- Consortium: 25 partners from 13 countries
  - National data centers, technology providers, research
- Objectives:
  - Cost-efficient and high-quality CDI
  - Meetings users' needs in flexible and sustainable way
  - Across geographical and disciplinary boundaries

http://www.eudat.eu

EUDAT Consortium

Data centers and Communities

EPOS: European Plate Observatory System

Research infrastructure and e-Science for data and observatories on earthquakes, volcanoes, surface dynamics and tectonics

- Distributed data sensors
- Large-scale statistics
- Metadata schema
- Reference architecture

EUDAT

CLARIN: Common Language Resources and Technology Infrastructure

CLARIN is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and usable

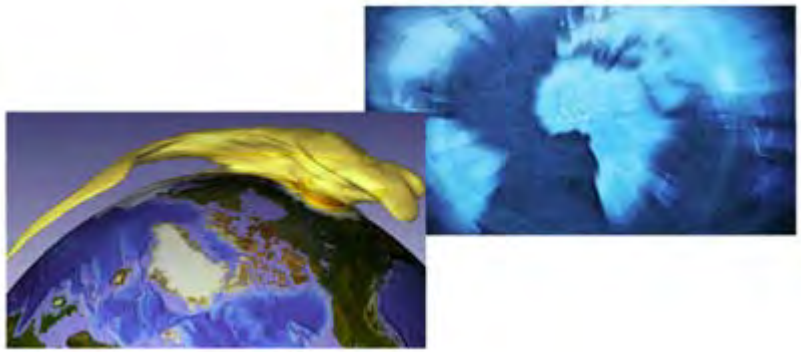
- Around 200 EU centres
- Require PIDs and metadata infrastructure
- ISOcat, SCHEMcat
- The Virtual Language Observatory
  - http://www.clarin.eu/vlo/

EUDAT

ENES: Service for Climate Modelling in Europe

ENES provides information and services to foster intricate simulations of the climate system using high-performance computers as well as the distribution and dissemination of data produced by such simulations

- About 20 EU centres
- Uses data infrastructure at the German climate centre
- Uses CIM data model
- Uses DOIs and EPIC handles
- Metadata schema based on ISO 11179



EUDAT

LifeWatch: Biodoversity Data and Observatories

LifeWatch will construct and bring into operation the facilities, hardware, software and governance structures for all aspects of biodiversity research: facilities for data generation and processing, data integration and interoperability; a network of observatories, virtual laboratories; a Service Centre supporting scientific and policy users

- Involving most “nature infrastructures”
- Interoperability requirements
- Distributed data sensors
- Metadata standardisation
- Common reference model

EUDAT

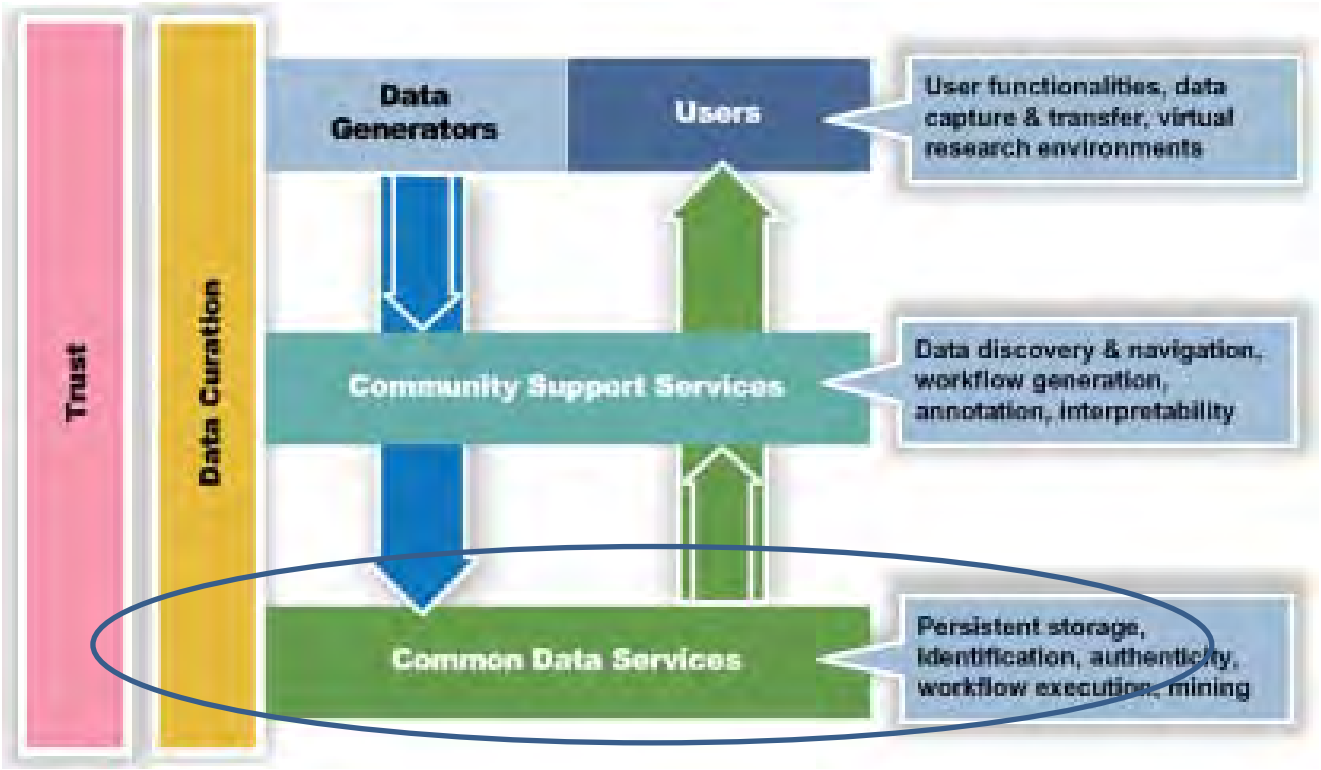
VPH: The Virtual Physiological Human

VPH aims to support and progress European research in biomedical modelling and simulation of the human body. This will improve our ability to predict, diagnose and treat disease, and have a dramatic effect on the future of healthcare, the pharmaceutical and medical device industries

- Pilot project with 5 hospitals
- Central datacentre
- Metadata aggregation
- DICOM, JPEG headers

EUDAT

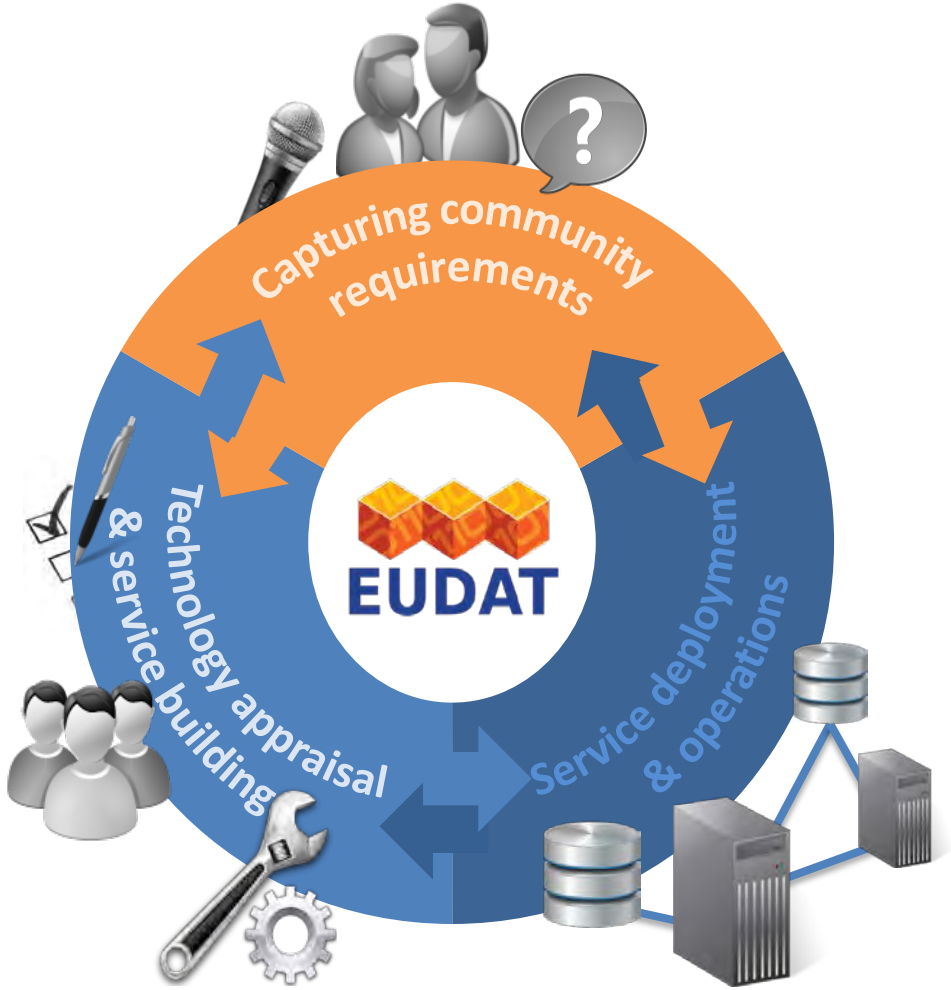
The CDI concept



Communities ↔ data centres

EUDAT

How do achieve this?



EUDAT

Building Blocks of the CDI

EUDAT Portal

Integrated APIs and harmonized access to EUDAT facilities

Metadata Catalog

Aggregated EUDAT metadata domain.
Data inventory

AAI

Network of trust among authentication and authorization actors

Data Staging

Dynamic replication to HPC workspace for processing

Safe Replication

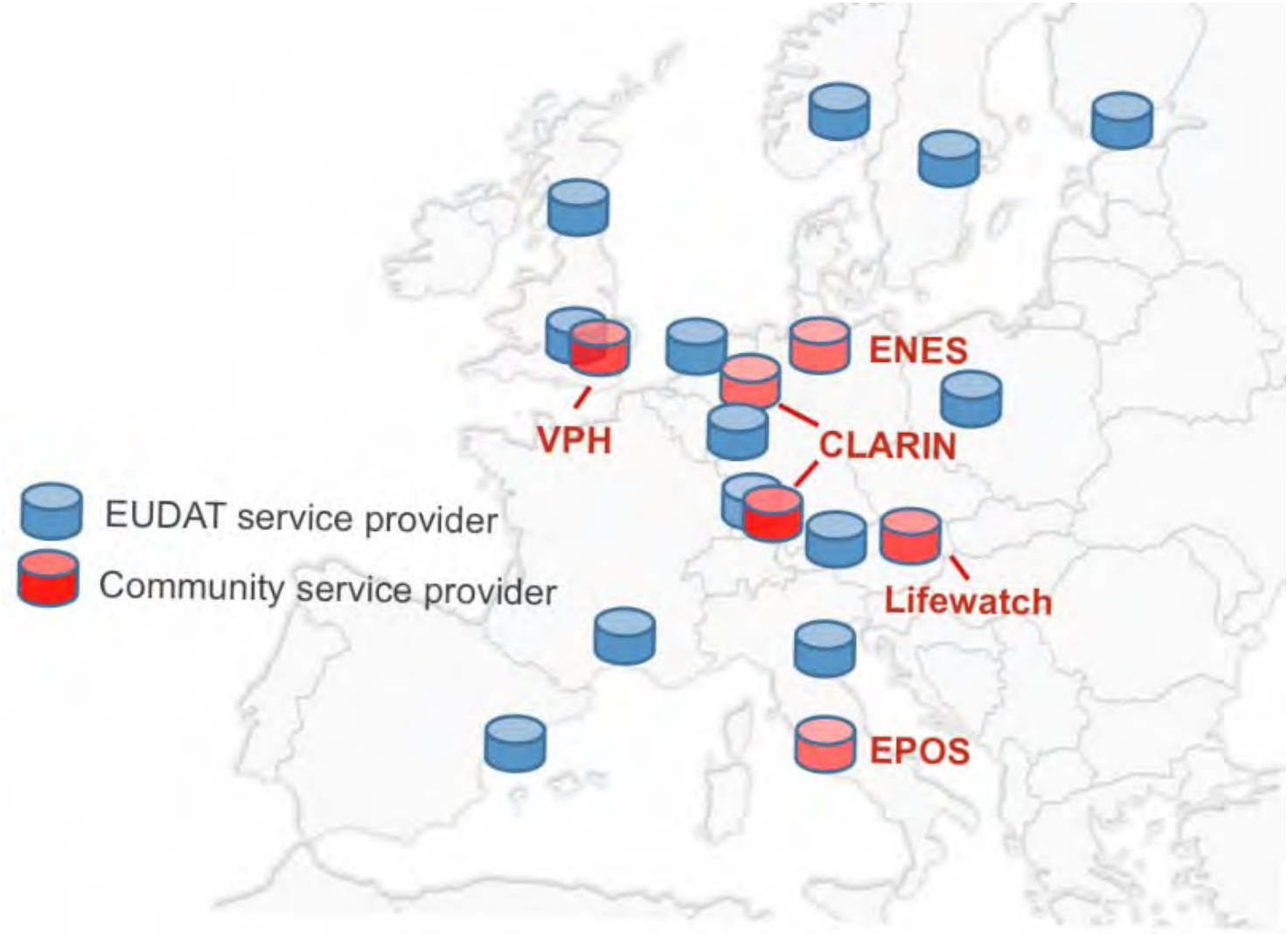
Data curation and access optimization

Simple Store

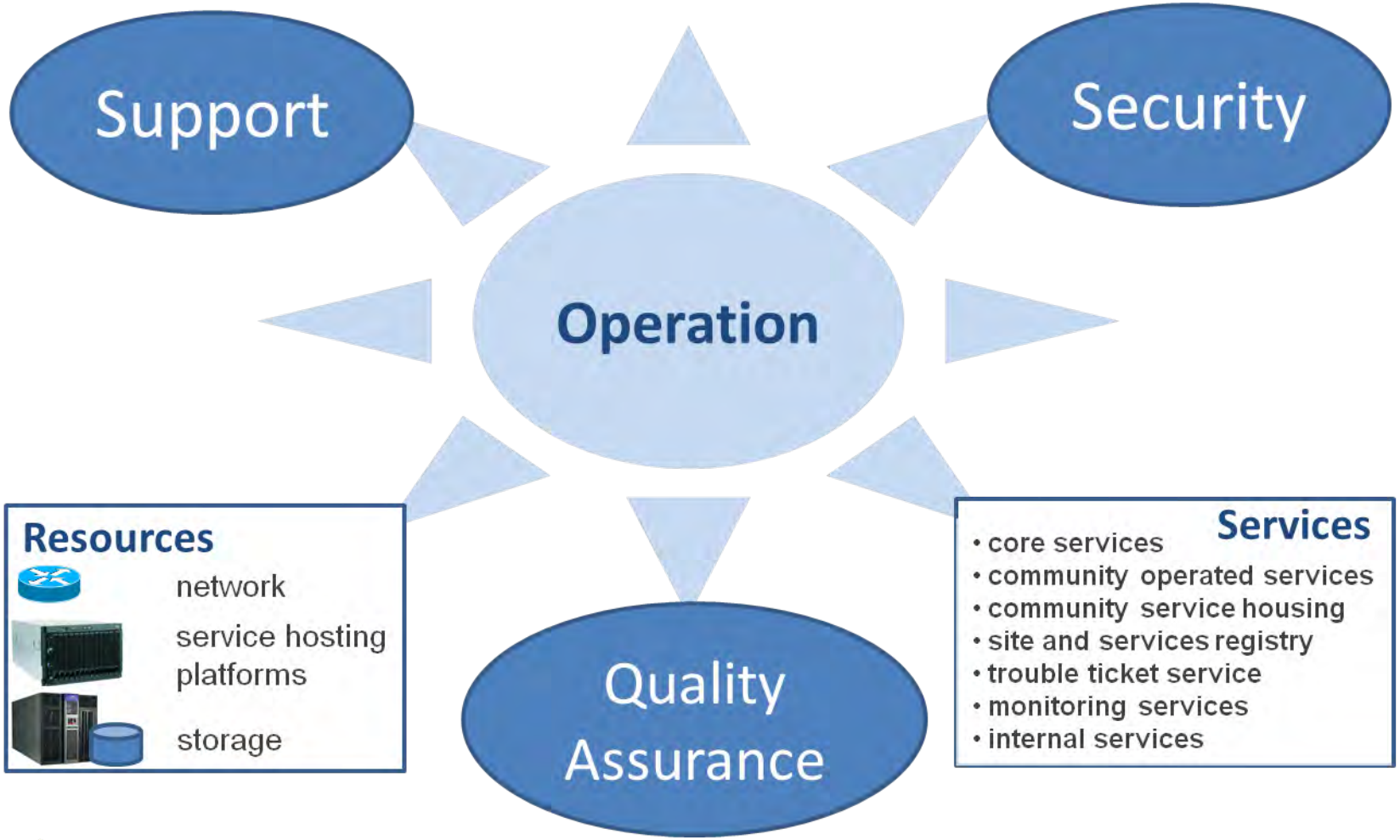
Researcher data store (simple upload, share and access)

EUDAT

INFRASTRUCTURE



OPERATION TEAM



How about data aquisition and licence agreements?

Some community examples: CLARIN

- **Type of Data**: text, audio, video, multimedia, etc.
- Aquisition, use, and re-use of data depends very much on the **source of the data**:
  - **Newspapers and publishing houses**: licences available to work with their data. Analysis belongs to the researcher (or his/her institution); but the underlying source remain property of the publishing house.
  - **Text, audio, or video interviews** created by researchers: exploitation right of the data (and maybe also copyright, depending on the situation) belongs to the researchers, but due to personality rights (of the persons interviewed), access to data can be restricted.
  - **Web-crawled text corpora**: permission of each website owner is required to distribute the texts
  - ➢ In most cases, copyright, personality right, and/or exploitation right are involved, i.e. hardly any linguistic data are completely freely available to the public. (except Wikipedia).
- **Licence agreements** (for annotated resources) cannot be standardised easily => uses of the resource needs the permission of the publishers or subjects.
- **Researchers** want to keep control of their data (for the reasons mentioned above). Also many researchers spent a lot of time for manually collecting data.
- **Trust** plays an important role – data owners need to keep control over their data.

Some community example: EPOS

- **Type of data**: from recording equipments (sensors, etc.) and lab experiments
- Acquisition, use, and re-use of data depends very much on the **source of the data AND "research groups"**:
  - **Seismologists and geodesists** obtaining their data from recording equipements installed in the field are generally very open to sharing the data because they also need the others' data => "community effort".
  - **Lab experiments** often require hosting costly and often unique machinery => attitude can be more "protective"
  - **Volcano scientists** – can be very protective of their data belonging to 'their' volcano.
- **Data restrictions** exists: some data is free and accessible in standard formats (e.g. data coming from permanent field equipments; other is restricted (e.g. data coming from labs). MoU exist between institutions to share data
- **Researchers** want to be innovative but also want to protect their work => because there is so much effort behind (i.e., for acquiring, quality checking, organising, etc.).
- EPOS seeks to address these issues and to combine the different data resources and create a single e-infrastructure where **data provenance, acknowledgment and right of use** are integral parts of the whole effort.

EUDAT

Some community example: LIFEWATCH

- **Type of data**: geospatial data, DNA sequences, coming from various disciplines.
- Generally it makes sense to make a difference between **different categories of data:**
  - Simple data strings (DNA sequences): open and free through the EBI portal
  - Signals (from sensors or Earth observation via satellites or plane): Feeling of ownership is stronger because of previous investments. Processed (filtered and transformed) data are also becoming more open and free available, but not in real time.
  - Processed data with often human interpretations (observations, monitoring): data mostly generated by academic and related research institutions (e.g. data about species presence and abundance, population densities, life stages of organisms, etc). As these data sets require curation, the ownership and owners responsibility must be clear.
- **Licences**: International agreements on in place for some data:
  - Open access (e.g. GBIF portal for sharing primary species presence data)
  - Free Access to metadata (e.g. LTER-Europe)
- Technical, semantic and legal **interoperability** is crucial for lifewatch

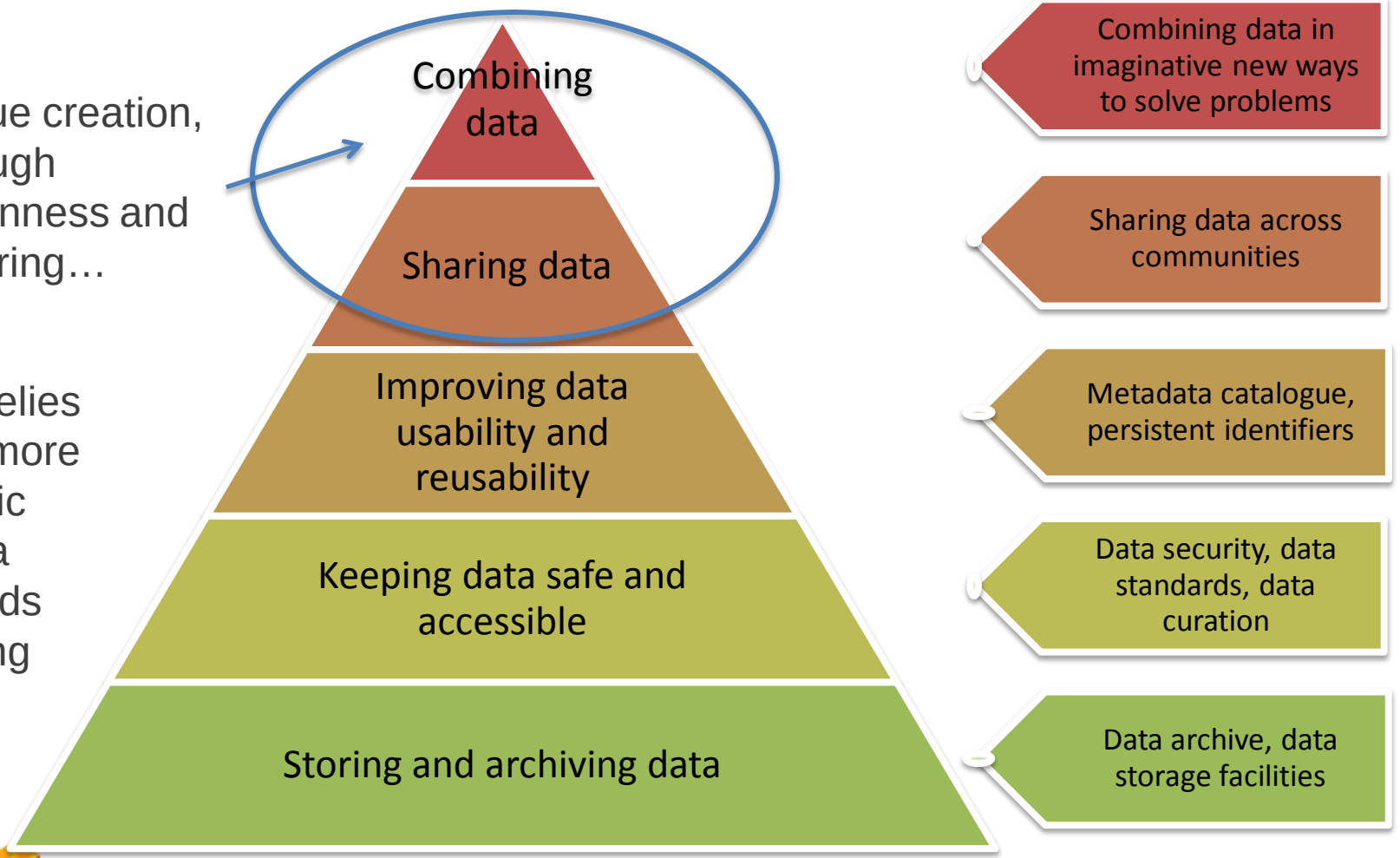
Some community example: ENES

- **Type of data**: climate modelling data
- Aquisition, use, and re-use of data depends very much on the **source of the data**
  - Data produced by the scientific community itself as climate model data:
    - Climate model data are available to the scientific community for academic use without restrictions but not anonymously.
    - About 2/3 of the modeling groups make their data available also for commerical applications according to the US model for data access.
  - Observational data like meteorological stations and satellites which are disseminated by agencies:
    - Data from observations and from satellites are available to the scientific community for academic use for free or on self-cost basis.
    - Commercial applications have to pay on an applications basis. The idea is that agencies will re-financed by selling data for commercial purposes.

Hierarchy of data needs

Value creation, though openness and sharing…

… relies on more basic data needs being met first



Principles – where we want to be

1. Data deposited with the EUDAT CDI will be preserved in perpetuity
2. Data are best curated in their own communities
3. Access to data in the EUDAT CDI is free at the point of use
4. EUDAT will operate as a federation of community-facing repositories and “back office” hosting providers
5. EUDAT services and infrastructure must be a suitable target for “TDR outsourcing” (cf. datasealofapproval.org)
6. EUDAT will not assert ownership of any data it holds

Welcome to the 1st EUDAT Conference!

Draft Programme

Monday 22nd October

Time	Title/Topic
9:00 - 17:00	**Project meetings**
9:00 - 17:00	**Training tutorials**

Tuesday 23rd October

Time	Title/Topic
8:00 - 9:30	*Registrations*
9:30 - 11:00	**Addressing the new data challenges – cross-disciplinary initiatives and open science** Data sharing and research excellence in astronomy and beyond: Synergies and tensions, Bernard F Schutz, MPG European data infrastructures in Horizon 2020, Kostas Glinos, EC EUDAT: Towards a collaborative data infrastructure, Kimmo Koski, CSC
11:00 - 11:30	*Coffee break*
11:30 - 12:45	**Developing common solutions through cluster initiatives** BioMedBridges: Constructing data and service bridges in the life sciences - Stephanie Suhr, EBI ENVRI: Operating an environmental RI - Wouter Los, UvA CRISP: Developing synergies in physics - Laurence Field, CERN DASISH: Weaving the SSH infrastructure - Daan Broeder /Johan Fihn
12:45 - 13:45	*Lunch*
13:45 - 15:45	**Parallel sessions I** 1. European e-Infrastructures collaboration 2. EUDAT services: Safe replication and data staging 3. EUDAT services: Metadata
15:45 - 16:15	*Coffee break*
16:15 - 18:00	**Lightning talks**
18:00 - 19:00	**Poster exhibition**
19:00 - 22:00	*Dinner*

Wednesday 24th October

Time	Title/Topic
8:30 - 9:30	*Registrations*
9:30 - 11:00	**Towards global data infrastructure components** An open architecture for managing information in the internet - Bob Kahn, CNRI Data - It's one world - Walter Stewart, research data co-ordinator, research data Canada iCORDI: A new platform for discussing global interoperability - Leif Laaksonen, CSC DAITF/DWF - Carlos Morais-Pires, EC
11:00 - 11:30	*Coffee break*
11:30 - 12:45	**Parallel sessions II** 1. Towards DAITF & DWF 2. EUDAT User Forum 3. Sustainability and funding models
12:45 - 13:15	**Wrap-up and Conclusions**
13:15 - 14:00	*Buffet lunch*

EUDAT

EUDAT

Contact us!

eudat-info@postit.csc.fi

www.eudat.eu

facebook.com/EUDAT

twitter.com/Eudat_eu

Project Coordinator: Kimmo Koski
kimmo.koski@csc.fi

Scientific Coordinator: Peter Wittenburg
peter.wittenburg@mpi.nl

Project Manager: Damien Lecarpentier
damien.lecarpentier@csc.fi

EUDAT Consortium

Data centers and Communities

EPOS: European Plate Observatory System

Research infrastructure and e-Science for data and observatories on earthquakes, volcanoes, surface dynamics and tectonics

- Distributed data sensors
- Large-scale statistics
- Metadata schema
- Reference architecture

EUDAT

CLARIN: Common Language Resources and Technology Infrastructure

CLARIN is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and usable

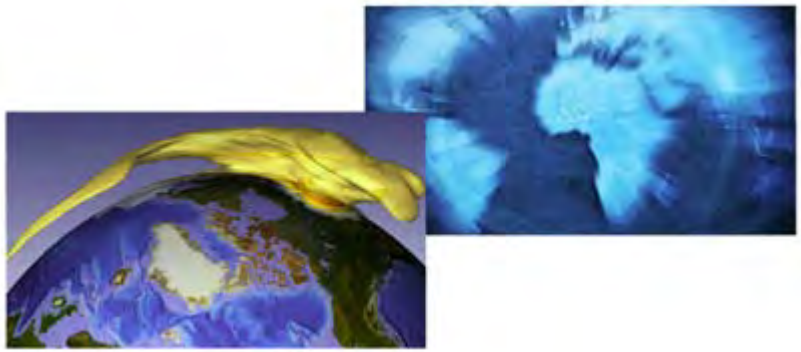
- Around 200 EU centres
- Require PIDs and metadata infrastructure
- ISOcat, SCHEMcat
- The Virtual Language Observatory
  - http://www.clarin.eu/vlo/

EUDAT

ENES: Service for Climate Modelling in Europe

ENES provides information and services to foster intricate simulations of the climate system using high-performance computers as well as the distribution and dissemination of data produced by such simulations

- About 20 EU centres
- Uses data infrastructure at the German climate centre
- Uses CIM data model
- Uses DOIs and EPIC handles
- Metadata schema based on ISO 11179



EUDAT

LifeWatch: Biodoversity Data and Observatories

LifeWatch will construct and bring into operation the facilities, hardware, software and governance structures for all aspects of biodiversity research: facilities for data generation and processing, data integration and interoperability; a network of observatories, virtual laboratories; a Service Centre supporting scientific and policy users

- Involving most “nature infrastructures”
- Interoperability requirements
- Distributed data sensors
- Metadata standardisation
- Common reference model

EUDAT

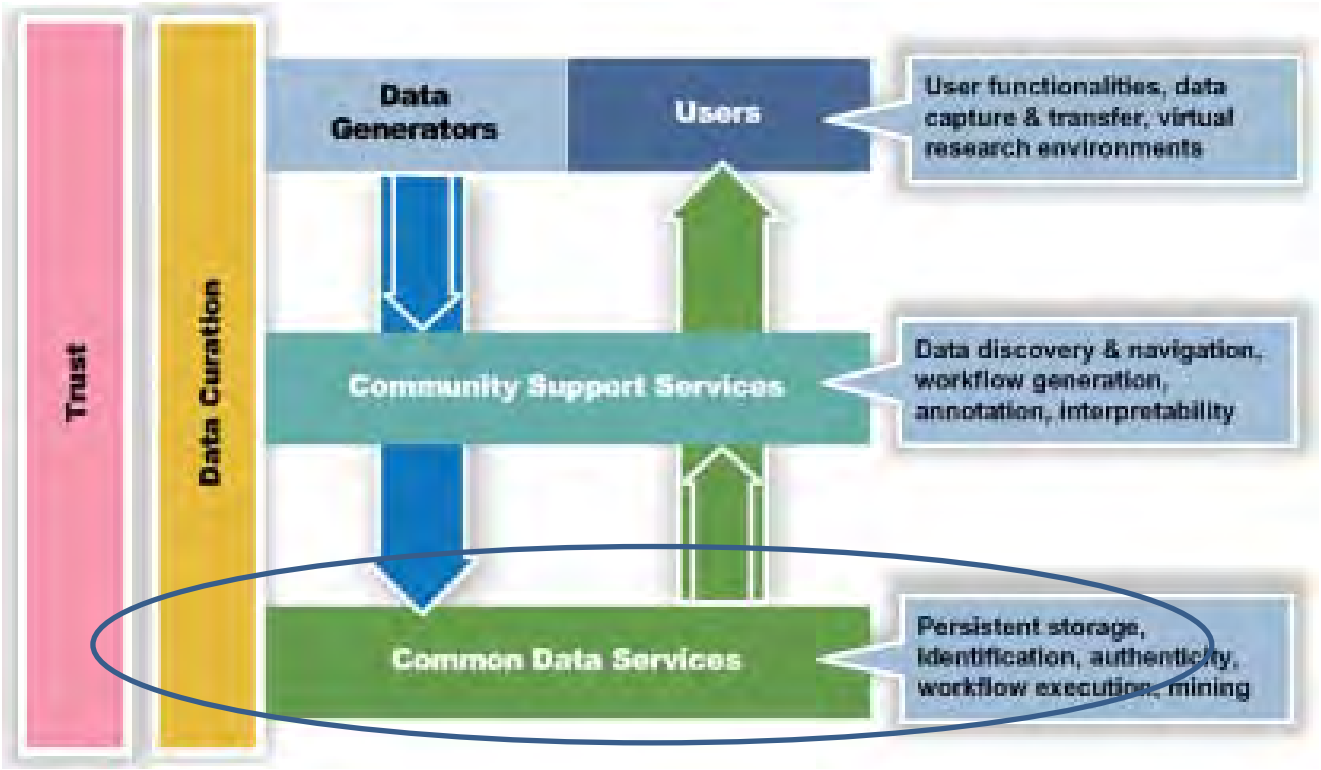
VPH: The Virtual Physiological Human

VPH aims to support and progress European research in biomedical modelling and simulation of the human body. This will improve our ability to predict, diagnose and treat disease, and have a dramatic effect on the future of healthcare, the pharmaceutical and medical device industries

- Pilot project with 5 hospitals
- Central datacentre
- Metadata aggregation
- DICOM, JPEG headers

EUDAT

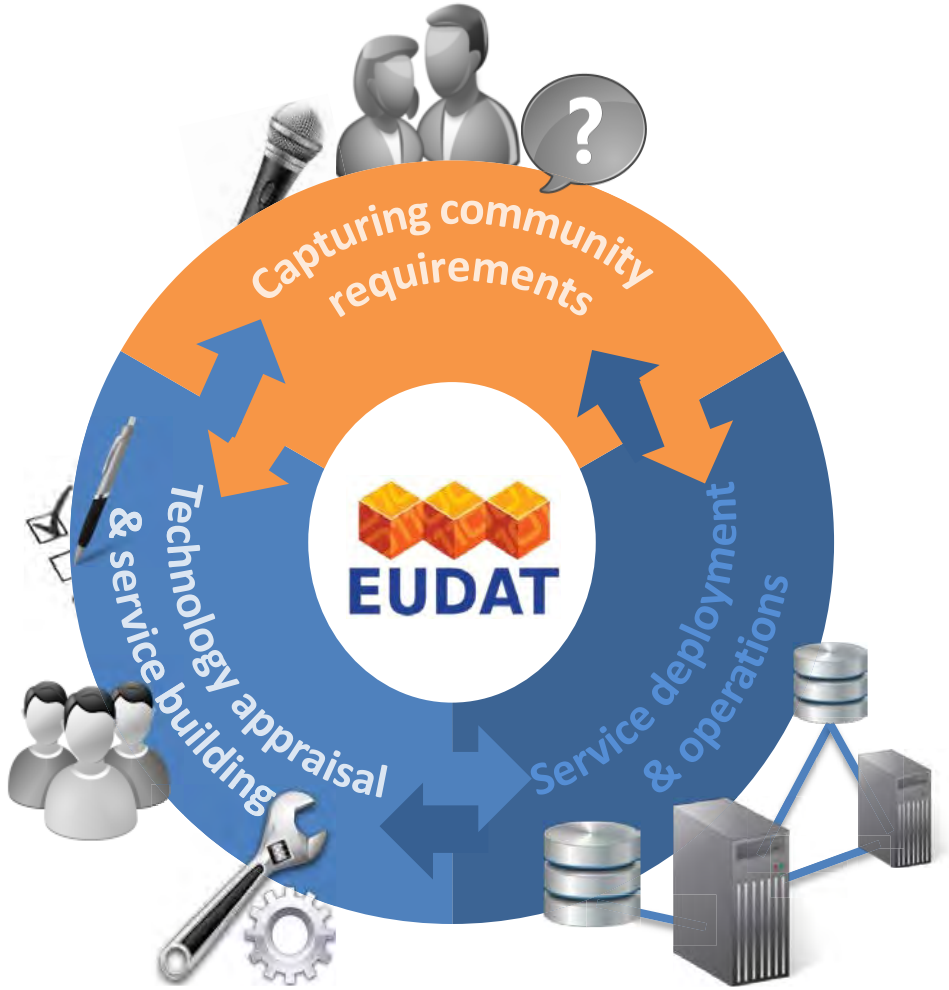
The CDI concept



Communities ↔ data centres

EUDAT

How do achieve this?



EUDAT

Building Blocks of the CDI

EUDAT Portal

Integrated APIs and harmonized access to EUDAT facilities

Metadata Catalog

Aggregated EUDAT metadata domain.
Data inventory

AAI

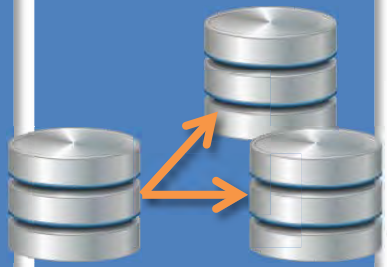
Network of trust among authentication and authorization actors

Data Staging

Dynamic replication to HPC workspace for processing

Safe Replication

Data curation and access optimization

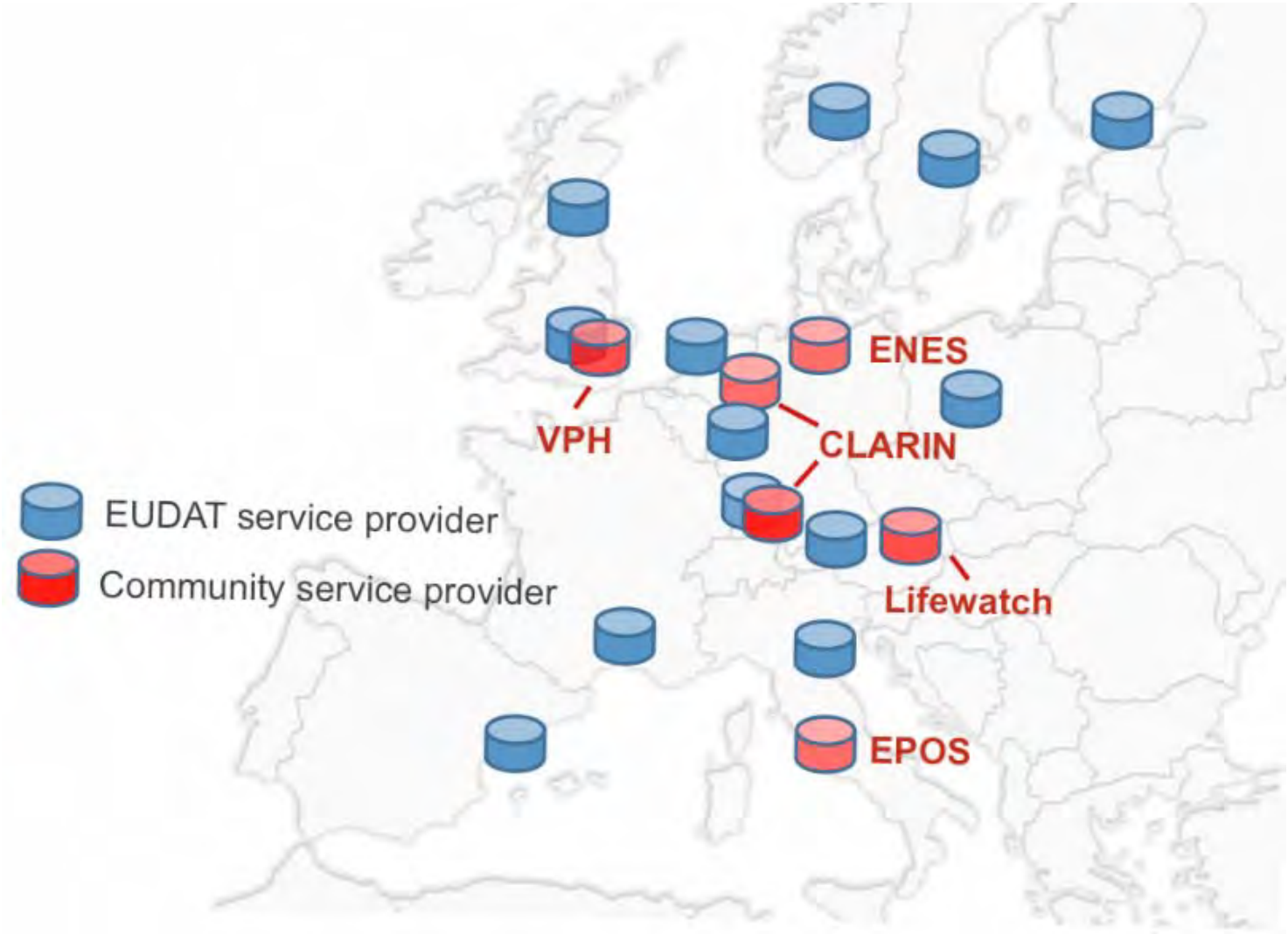


Simple Store

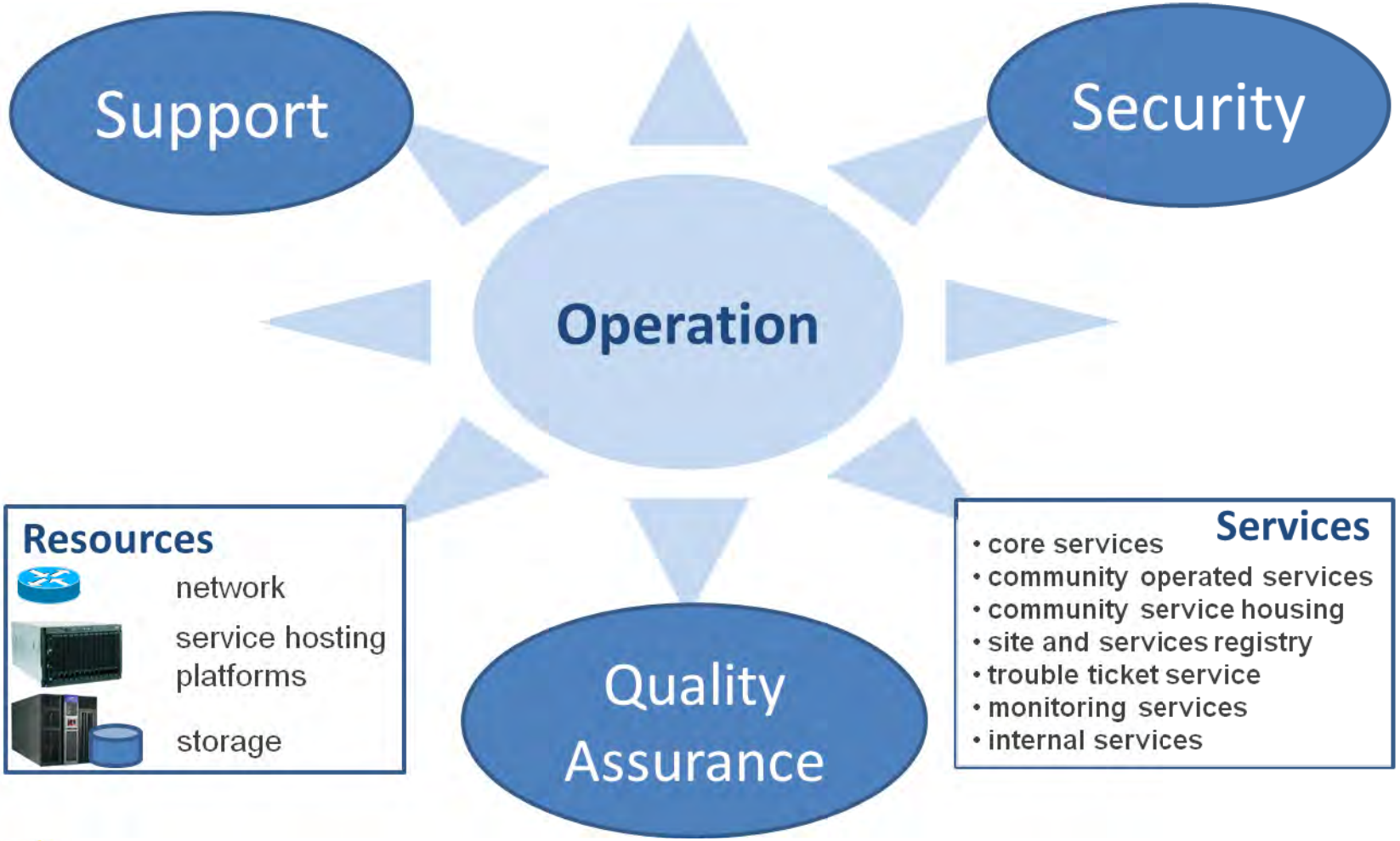
Researcher data store (simple upload, share and access)

EUDAT

INFRASTRUCTURE



OPERATION TEAM



How about data aquisition and licence agreements?

Some community examples: CLARIN

- **Type of Data**: text, audio, video, multimedia, etc.
- Aquisition, use, and re-use of data depends very much on the **source of the data**:
  - **Newspapers and publishing houses**: licences available to work with their data. Analysis belongs to the researcher (or his/her institution); but the underlying source remain property of the publishing house.
  - **Text, audio, or video interviews** created by researchers: exploitation right of the data (and maybe also copyright, depending on the situation) belongs to the researchers, but due to personality rights (of the persons interviewed), access to data can be restricted.
  - **Web-crawled text corpora**: permission of each website owner is required to distribute the texts
  - ➢ In most cases, copyright, personality right, and/or exploitation right are involved, i.e. hardly any linguistic data are completely freely available to the public. (except Wikipedia).
- **Licence agreements** (for annotated resources) cannot be standardised easily => uses of the resource needs the permission of the publishers or subjects.
- **Researchers** want to keep control of their data (for the reasons mentioned above). Also many researchers spent a lot of time for manually collecting data.
- **Trust** plays an important role – data owners need to keep control over their data.

Some community example: EPOS

- **Type of data**: from recording equipments (sensors, etc.) and lab experiments
- Acquisition, use, and re-use of data depends very much on the **source of the data AND "research groups"**:
  - **Seismologists and geodesists** obtaining their data from recording equipements installed in the field are generally very open to sharing the data because they also need the others' data => "community effort".
  - **Lab experiments** often require hosting costly and often unique machinery => attitude can be more "protective"
  - **Volcano scientists** – can be very protective of their data belonging to 'their' volcano.
- **Data restrictions** exists: some data is free and accessible in standard formats (e.g. data coming from permanent field equipments; other is restricted (e.g. data coming from labs). MoU exist between institutions to share data
- **Researchers** want to be innovative but also want to protect their work => because there is so much effort behind (i.e., for acquiring, quality checking, organising, etc.).
- EPOS seeks to address these issues and to combine the different data resources and create a single e-infrastructure where **data provenance, acknowledgment and right of use** are integral parts of the whole effort.

Some community example: LIFEWATCH

- **Type of data**: geospatial data, DNA sequences, coming from various disciplines.
- Generally it makes sense to make a difference between **different categories of data:**
  - Simple data strings (DNA sequences): open and free through the EBI portal
  - Signals (from sensors or Earth observation via satellites or plane): Feeling of ownership is stronger because of previous investments. Processed (filtered and transformed) data are also becoming more open and free available, but not in real time.
  - Processed data with often human interpretations (observations, monitoring): data mostly generated by academic and related research institutions (e.g. data about species presence and abundance, population densities, life stages of organisms, etc). As these data sets require curation, the ownership and owners responsibility must be clear.
- **Licences**: International agreements on in place for some data:
  - Open access (e.g. GBIF portal for sharing primary species presence data)
  - Free Access to metadata (e.g. LTER-Europe)
- Technical, semantic and legal **interoperability** is crucial for lifewatch

EUDAT

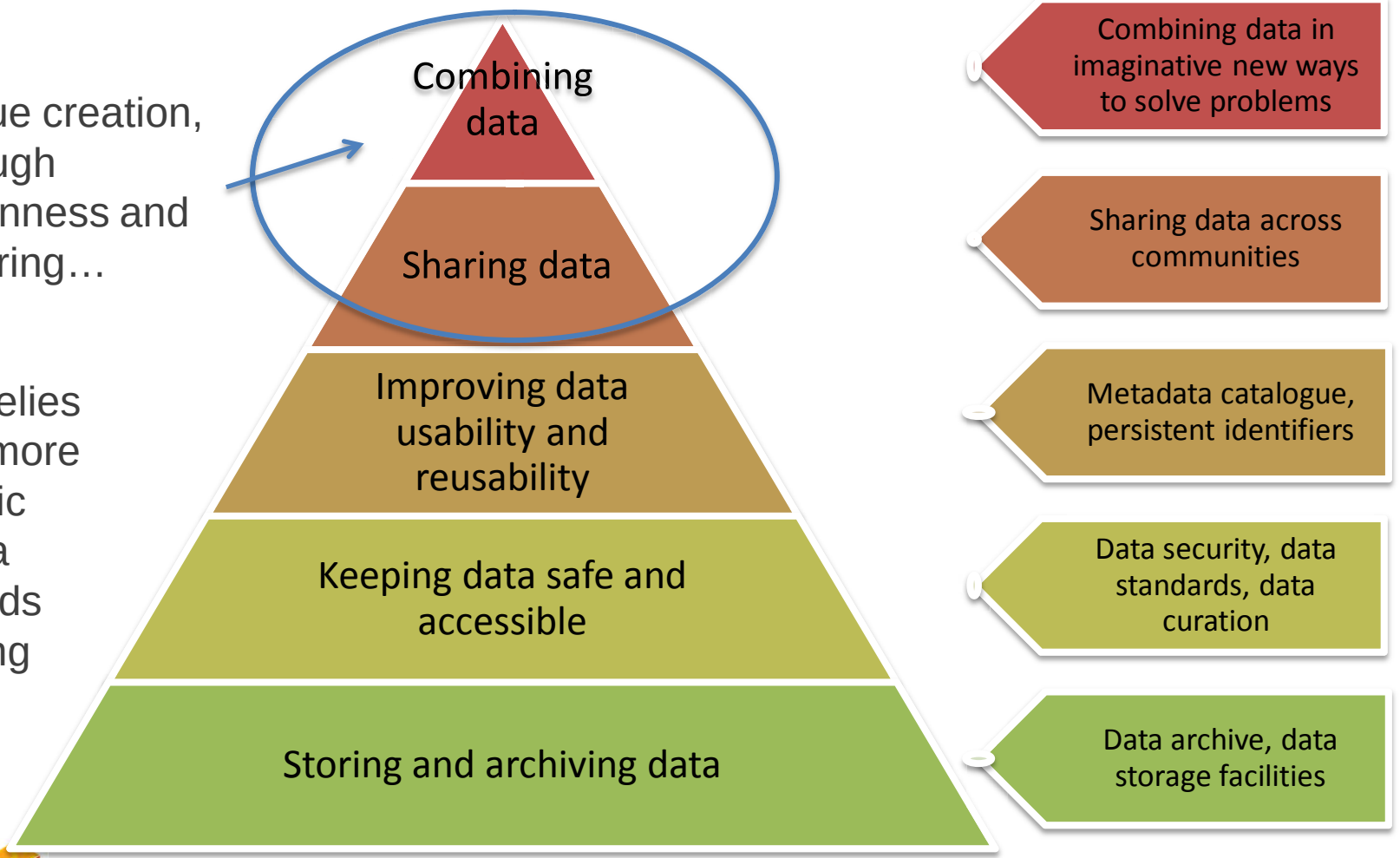
Some community example: ENES

- **Type of data**: climate modelling data
- Aquisition, use, and re-use of data depends very much on the **source of the data**
  - Data produced by the scientific community itself as climate model data:
    - Climate model data are available to the scientific community for academic use without restrictions but not anonymously.
    - About 2/3 of the modeling groups make their data available also for commerical applications according to the US model for data access.
  - Observational data like meteorological stations and satellites which are disseminated by agencies:
    - Data from observations and from satellites are available to the scientific community for academic use for free or on self-cost basis.
    - Commercial applications have to pay on an applications basis. The idea is that agencies will re-financed by selling data for commercial purposes.

Hierarchy of data needs

Value creation, though openness and sharing…

… relies on more basic data needs being met first



Principles – where we want to be

1. Data deposited with the EUDAT CDI will be preserved in perpetuity
2. Data are best curated in their own communities
3. Access to data in the EUDAT CDI is free at the point of use
4. EUDAT will operate as a federation of community-facing repositories and “back office” hosting providers
5. EUDAT services and infrastructure must be a suitable target for “TDR outsourcing” (cf. datasealofapproval.org)
6. EUDAT will not assert ownership of any data it holds

Welcome to the 1st EUDAT Conference!

Draft Programme

Monday 22nd October

Time	Title/Topic
9:00 - 17:00	**Project meetings**
9:00 - 17:00	**Training tutorials**

Tuesday 23rd October

Time	Title/Topic
8:00 - 9:30	*Registrations*
9:30 - 11:00	**Addressing the new data challenges – cross-disciplinary initiatives and open science** Data sharing and research excellence in astronomy and beyond: Synergies and tensions, Bernard F Schutz, MPG European data infrastructures in Horizon 2020, Kostas Glinos, EC EUDAT: Towards a collaborative data infrastructure, Kimmo Koski, CSC
11:00 - 11:30	*Coffee break*
11:30 - 12:45	**Developing common solutions through cluster initiatives** BioMedBridges: Constructing data and service bridges in the life sciences - Stephanie Suhr, EBI ENVRI: Operating an environmental RI - Wouter Los, UvA CRISP: Developing synergies in physics - Laurence Field, CERN DASISH: Weaving the SSH infrastructure - Daan Broeder /Johan Fihn
12:45 - 13:45	*Lunch*
13:45 - 15:45	**Parallel sessions I** 1. European e-Infrastructures collaboration 2. EUDAT services: Safe replication and data staging 3. EUDAT services: Metadata
15:45 - 16:15	*Coffee break*
16:15 - 18:00	**Lightning talks**
18:00 - 19:00	**Poster exhibition**
19:00 - 22:00	*Dinner*

Wednesday 24th October

Time	Title/Topic
8:30 - 9:30	*Registrations*
9:30 - 11:00	**Towards global data infrastructure components** An open architecture for managing information in the internet - Bob Kahn, CNRI Data - It's one world - Walter Stewart, research data co-ordinator, research data Canada iCORDI: A new platform for discussing global interoperability - Leif Laaksonen, CSC DAITF/DWF - Carlos Morais-Pires, EC
11:00 - 11:30	*Coffee break*
11:30 - 12:45	**Parallel sessions II** 1. Towards DAITF & DWF 2. EUDAT User Forum 3. Sustainability and funding models
12:45 - 13:15	**Wrap-up and Conclusions**
13:15 - 14:00	*Buffet lunch*

EUDAT

EUDAT

Contact us!

eudat-info@postit.csc.fi

www.eudat.eu

facebook.com/EUDAT

twitter.com/Eudat_eu

Project Coordinator: Kimmo Koski
kimmo.koski@csc.fi

Scientific Coordinator: Peter Wittenburg
peter.wittenburg@mpi.nl

Project Manager: Damien Lecarpentier
damien.lecarpentier@csc.fi