Towards an Open Science Commons

Tiziana Ferrari

EGI.eu Technical Director

- The European Grid Infrastructure today
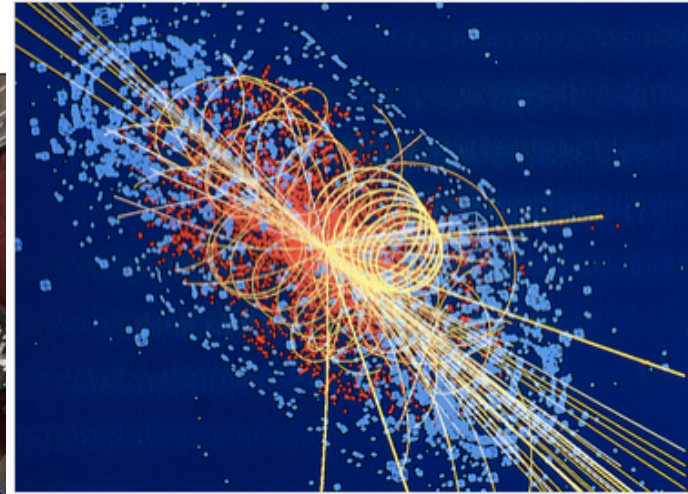
- Medium-term plans

- Towards 2020

EGI today

- Governance, status, in other words the power of the federated model

# From HEP to international multidisciplinary open science

Peter Higgs stands on the cavern floor at CMS, with the detector open for maintenance behind him (Image: Maximilien Brice/CERN)

François Englert (left) and Peter Higgs at CERN on 4 July 2012, on the occasion of the announcement of the discovery of a Higgs boson by the ATLAS and CMS experiments (Image: Maximilien Brice/CERN)

EGI and its participants - 2015

- ## 25 European participants: 23 NGIs and 2 EIROs (CERN, EMBL-EBI)

  - Opening membership to research communities

- ## Affiliation programme

  - lower barriers of entry to widening countries

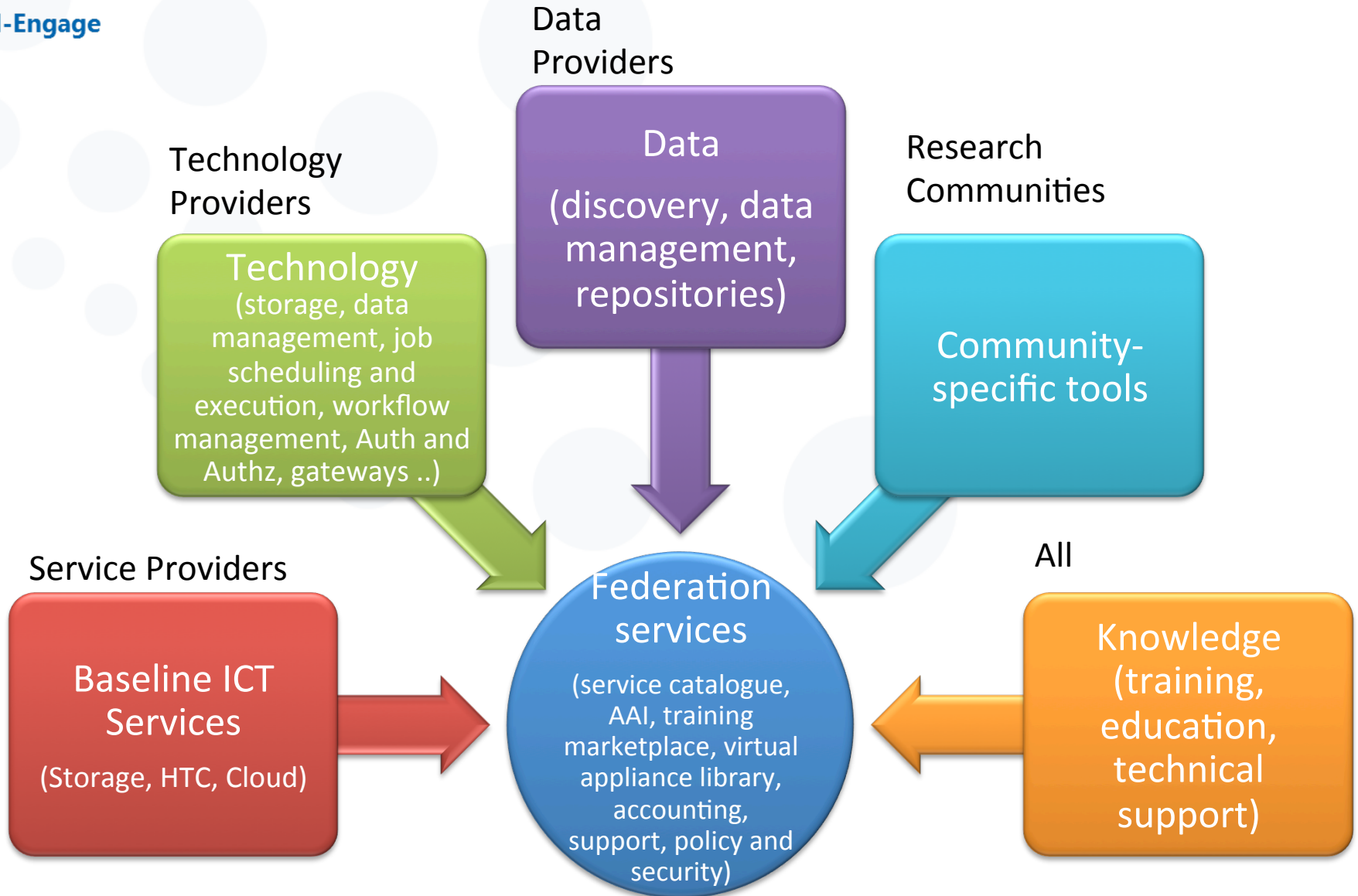EGI.eu in Amsterdam
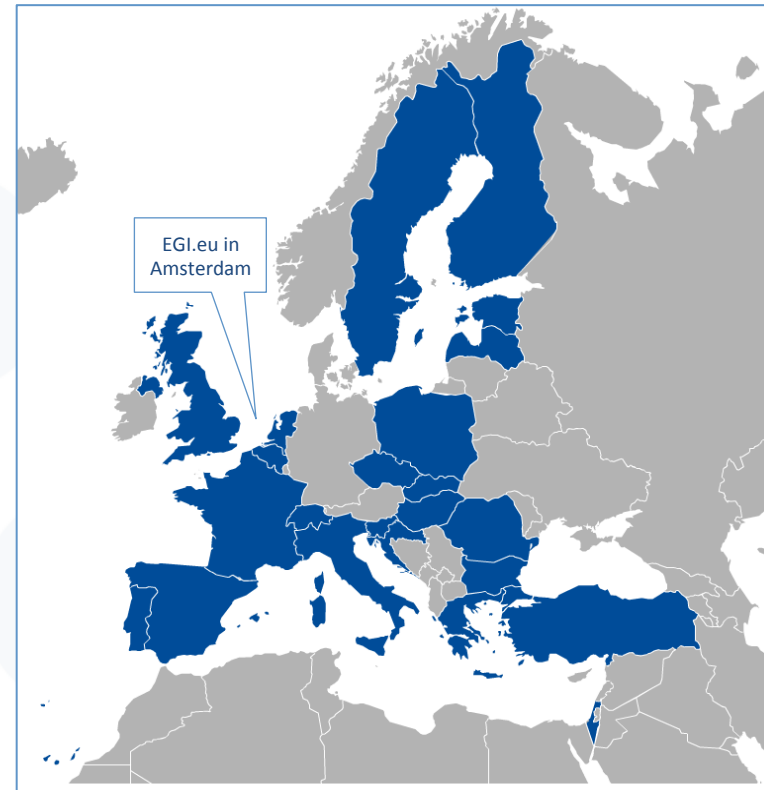
Participants
CERN, EMBL-EBI, Belgium, Bulgaria, Croatia, Czech Republic, Estonia, Finland, France, Greece, Hungary, Israel, Italy, FYR of Macedonia, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Switzerland, Sweden, Turkey, UK

Under discussion
Armenia, Austria, Belarus, Germany, Denmark, Moldova, Norway, Russia, Ukraine

- High-Throughput Data analysis

- Federated Cloud

- (New) Federated Open Data Processing through a

  – Open data programme

  – Virtual Research Environment programme

- Federated Operations

- Community driven Innovation and Support

- Policy Advice

Federating open science

Data Providers

Data
(discovery, data management, repositories)

Technology Providers

Technology
(storage, data management, job scheduling and execution, workflow management, Auth and Authz, gateways ..)

Research Communities

Community-specific tools

Service Providers

Baseline ICT Services
(Storage, HTC, Cloud)

Federation services
(service catalogue, AAI, training marketplace, virtual appliance library, accounting, support, policy and security)

All

Knowledge
(training, education, technical support)

Science is inherently distributed

- **Discoverability** of services and knowledge
- **Portability and open standards**
  - data, applications, software
- **Sharing** and **openness**
- Common **access policies, security**
- One **accounting** infrastructure
- One **support** infrastructure
- **Single sign on**
- Federated service management
- Aggregation of **demand and offer**

Federating e-infrastructures and data 1/2

Billions of hours (wall clock hours, HEPSPEC06 normalized)
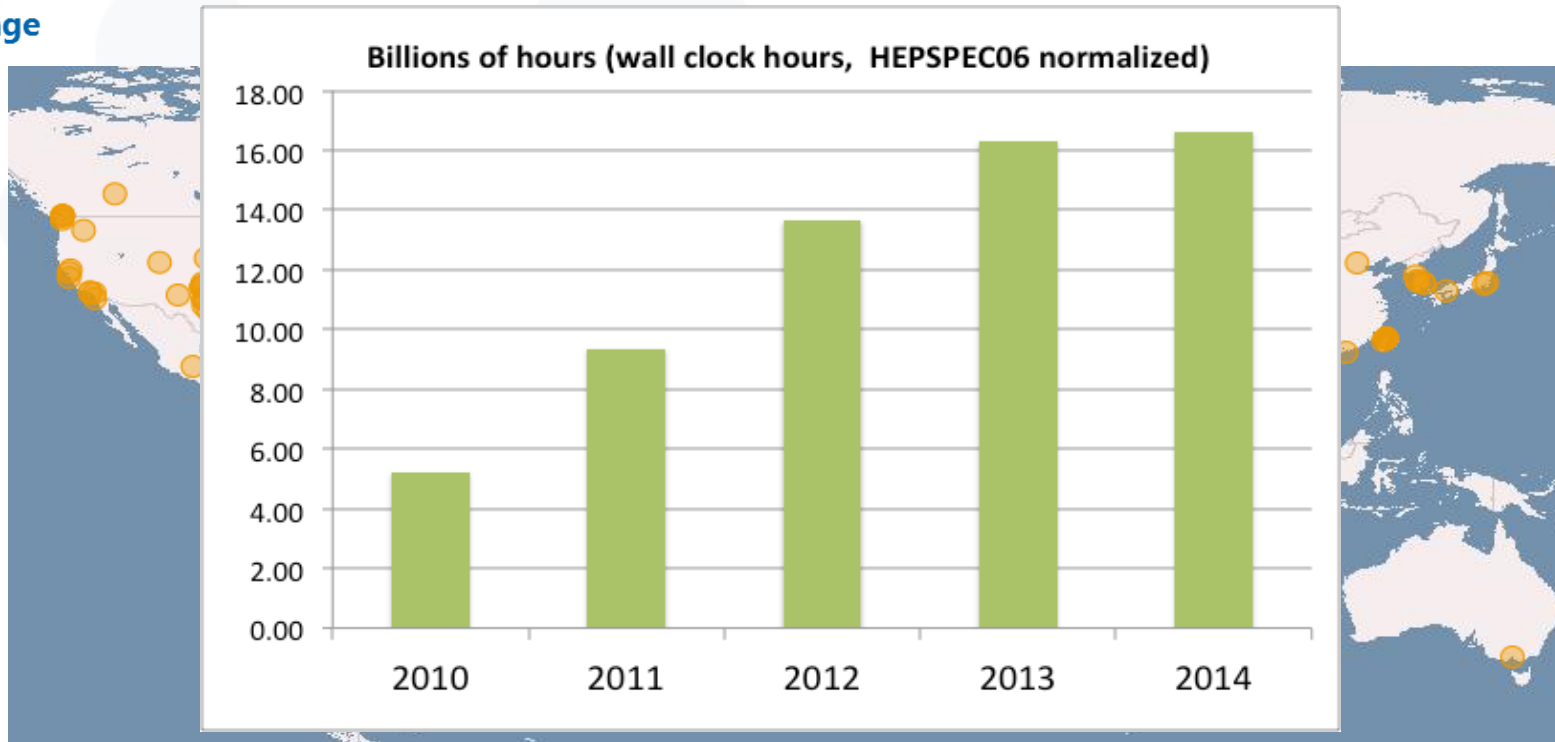
- Distributed, federated storage, HTC and cloud facilities
- Virtual Research Environments
- > 200 registered user research projects

- 340 resource centres in 54 countries
- 550,000 logical CPU cores
- >290 PB disk, 180 PB tape
- > 99.6% reliability

Federating e-infrastructures and data 2/2

EGI Council members
Integrated Infrastructures
Peer Infrastructures

Academia Sinica
Grid Computing

CLAF

→ **More than 6,000 jobs/year to OSG**
→ **More than 68,000 jobs/year in IDGF**
→ **840 M CPU hours/year in Asia Pacific**

Achievements

- Operations
- Platforms
- User outreach
- Strategy and policy

EGI today

Get infrastructure services

Resource allocation for national and international resources

- e-GRANT
  - Pooling of distributed infrastructure resources (HTC and cloud)
  - Matchmaking demand ←→ offer
  - Allocation
  - SLA negotiation (user community ←→ EGI.eu)

- Monitoring of service level targets
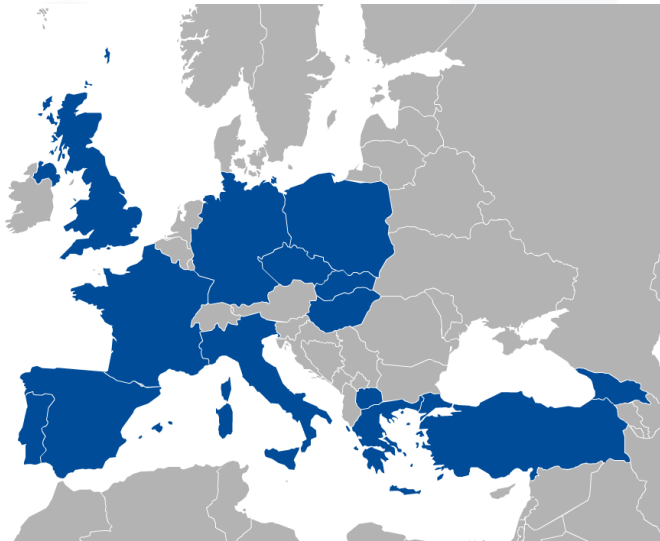
EGI Federated cloud

- 14 countries, 21 providers
- 9 M CPU hours in 12 months
- Hybrid federation
  - Public clouds (open to any research community, based on open cloud standards for portability of applications and data)
  - Community clouds (for selected list of VOs, looser federation profile based on a subset of federation tools)
  - Bringing cloud services next to big data
  - Federated AAI, accounting, discovery and monitoring

Usage Models

FedCloud extends EGI computing beyond HTC with new usage models:

- **Service Hosting**
  - Long-running services (e.g. web, database or application servers)
- **Compute and data intensive workloads**
  - Batch and interactive (e.g. IPython, R, matlab) with scalable and customized environments not limited to the traditional job model
- **Datasets repository**
  - Store and manage large datasets for your applications
- **Disposable and testing environments**
  - Host training events, test new developments and applications without overhead

VM Management

On demand compute to run any kind of workloads on virtual machines

Site A

VM   VM

Site B

VM   VM
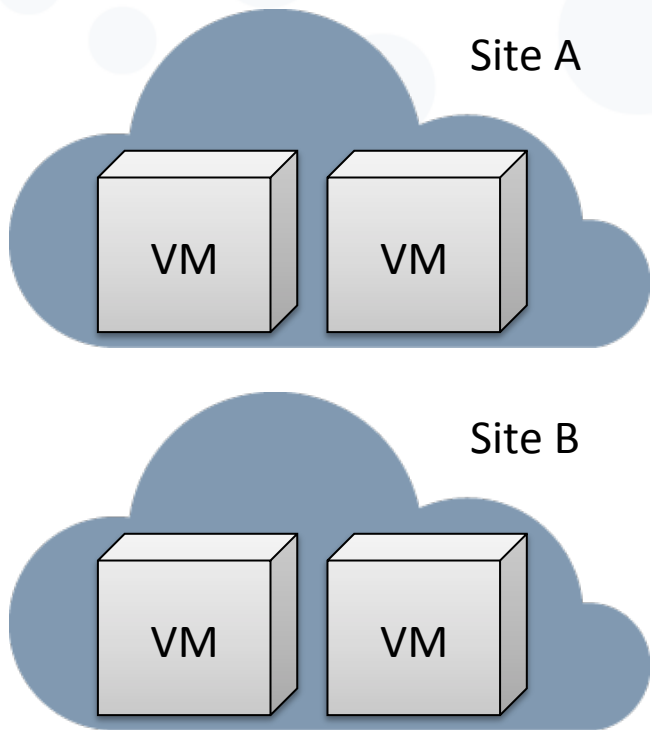
Easy provisioning

- OCCI API across the whole infrastructure
- VMs start immediately
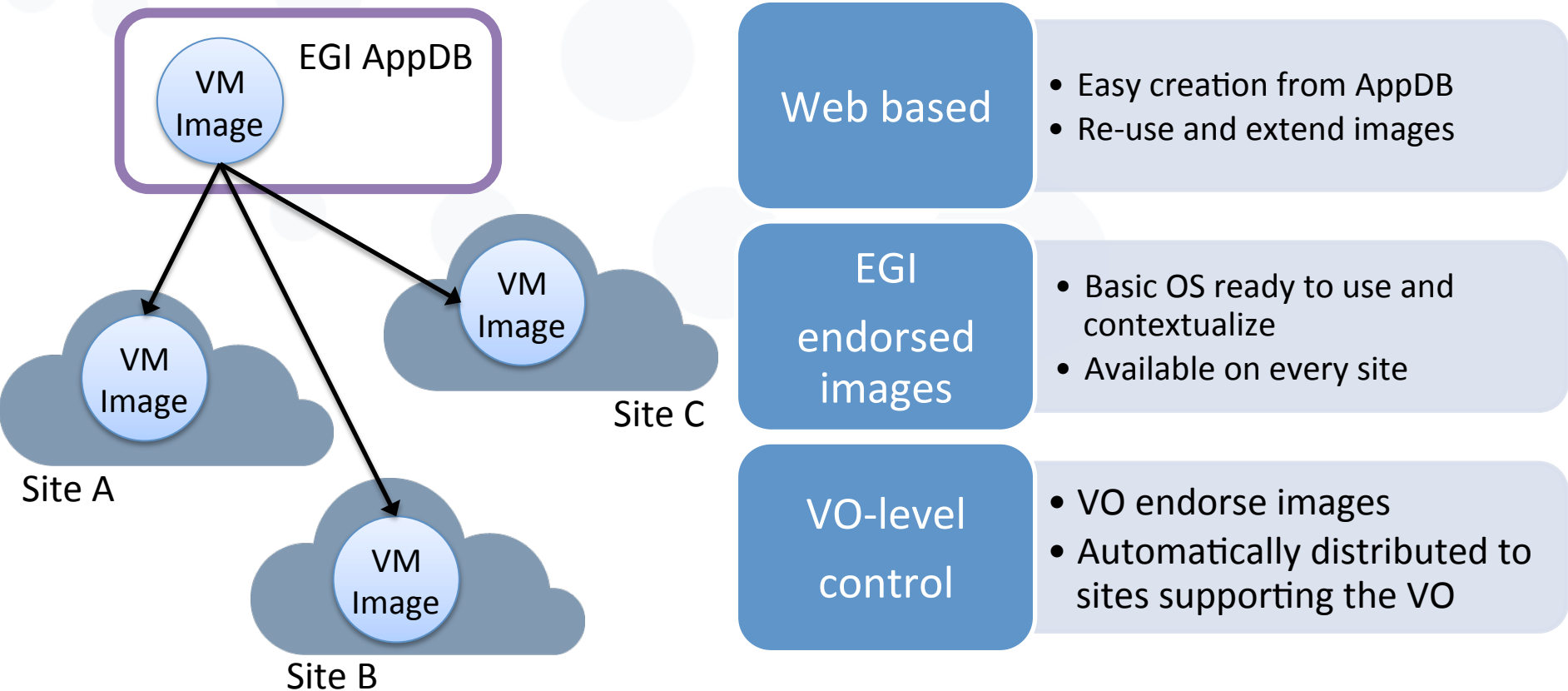- Ruby and Java clients

Customize

- Select your OS
- *root* access
- Contextualization

Scale to your needs

- Select VM size (cores, RAM)
- Create and destroy VMs as needed

VM Image Management

Automatic and secure distribution of endorsed VM images for Virtual Organisations

EGI AppDB

VM Image

VM Image

Site C

VM Image

Site A

VM Image

Site B

Web based
- Easy creation from AppDB
- Re-use and extend images

EGI endorsed images
- Basic OS ready to use and contextualize
- Available on every site

VO-level control
- VO endorse images
- Automatically distributed to sites supporting the VO

Storage management

• Block Storage

- **Persistent** Block Level Storage to **attach** to VMs
- Manage with **OCCI**
  - use as any other block device from VMs
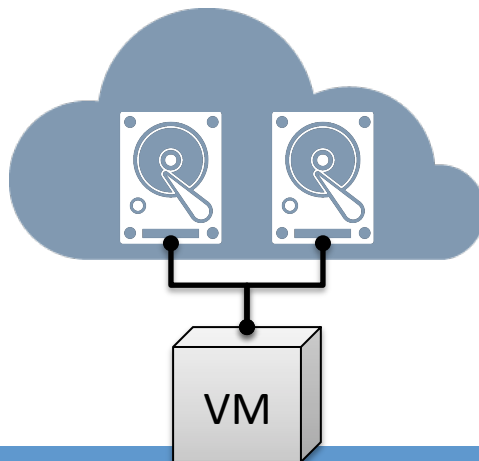
• Object Storage

- **Data storage** accessible fro **anywhere** at **any time**
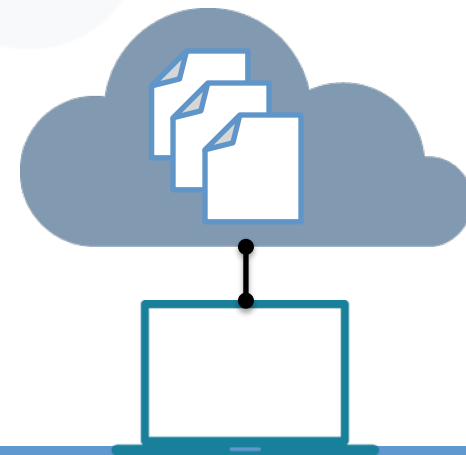- **Sharing/Serve** data
- **CDMI** as standard API
  - No POSIX access!

- A community driven central service that stores and provides:
  - ❑ **software solutions** (of *any* form, native software and/or virtual appliances), originated from almost every scientific area/discipline
  - ❑ reference of scientific **datasets** (under development)
  - ❑ the **programmers** and **scientists** responsible for them
  - ❑ the **publications** derived from the registered items (SW, VA & datasets)

Acts as **distribution** medium, for:
  - ❑ native **software packages** (RPMs, DEBs, TARs) exposed through the UMD Community Repository to the public
  - ❑ **virtual machine images** by using the HEPiX virtualization vmcatcher/vmcaster technology

Value proposition

- Foster the sharing of scientific codes, applications tools → application software library, communities of scientific codes

- Share and manage virtual appliances → reuse

- Link virtual appliances to datasets and publications → reproducibility of science

Cloud marketplace

- Holds, manages and populates "**Virtual Appliances**"
  - Virtual Appliance => bundle of one or more Virtual Machine Images ready to be used either on a cloud infrastructure (i.e. EGI FedCloud) or even to be downloaded for personal use

- Ability to the user to create "**Software Appliances**"
  - Software Appliance: a pair of a Virtual Appliance and a contextualization script which can be used for adhoc software installations at the Virtual Appliance during its boot time i.e. a base Virtual Appliance of a Centos6 and a contextualization script that downloads and installs the latest version of octave, on boot time.

- **VO or Project Managers**: able to **select** which of the registered Virtual Appliances, should be pushed to the sites of their responsibility.
  - The system takes the responsibility for **distributing** the "**chosen**" ones to the respective Resource providers/sites.
  - AppDB **gets feedback from the sites** that the Virtual Appliances **are indeed at the site side** and are ready to be used by the users.

The marketplace in action

- Search for vAppliances
- Get **availability** & **usage** details
- Start/Stop/…. an image

perform brokering cmds

Next step:
perform brokering through the AppDB portal

users

AppDB portal

Remote Sites

cloud m/w

local info system

> vmcatcher

push info about the VA **availability** and **usage**

Information System (Top-BDII for FedCloud sites)

subscribe and/or get any update

AppDB Image List Store

generate image_list

AppDB

Cloud Marketplace

- (Re)publish VO or Project's accepted list of Virtual Appliances
- Select/update/exclude vAppliances

- Publish new vApliance version
- Insert/update/remove image(s)
- Create a new vApliance version
- Register a vAppliance

AppDB portal

VO or Project Manager

AppDB portal

> vmcaster

submitter

- 26 communities
  - Biological sciences
  - Physical sciences
  - Earth sciences
- 59 use cases currently supported, 5 from commercial organisations
- 700,000 VMs instantiated

Strategy, Policy, Business Development

- The Open Science Commons

- Which business model? Free vs pay-for-use
  - Pay 30 providers across 12 countries publishing pricing information (~10 ready/able to sell)
  - Emerging business models
  - Tools adapted (GOCDB, AppDB, e-GRANT), including GUI
  - Final Report

Impact
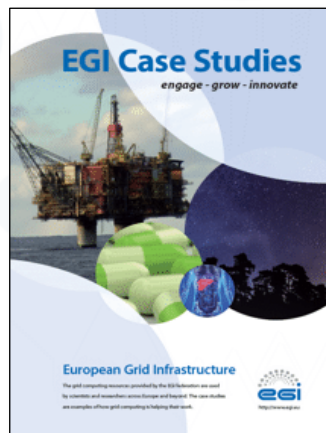
EGI case studies

3,600 service end-points, 47 UMD releases, 38,000 users

Increasing use of new disciplines

- 220 research projects, 76 new
- Astronomy and astro-particle Physics, Structural biology, Hydrology and climate, Medical and Health Sciences
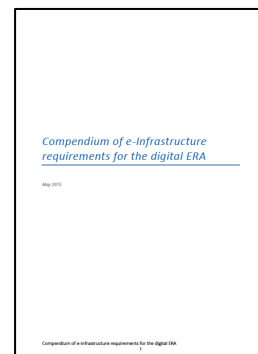
Better services for the long tail

- 46% of the new users)

Support to Research Infrastructures

- BBMRI, CTA
- Testing: EISCAT-3D, ELIXIR, ELI-NP, LifeWatch, LOFAR, KM3NeT

2,400 Peer-reviewed papers, 620 new registered applications

Compendium of RI requirements

EGI today

Achievements

The big shifts through EGI-Engage and sister projects

- AARC
- INDIGO Datacloud
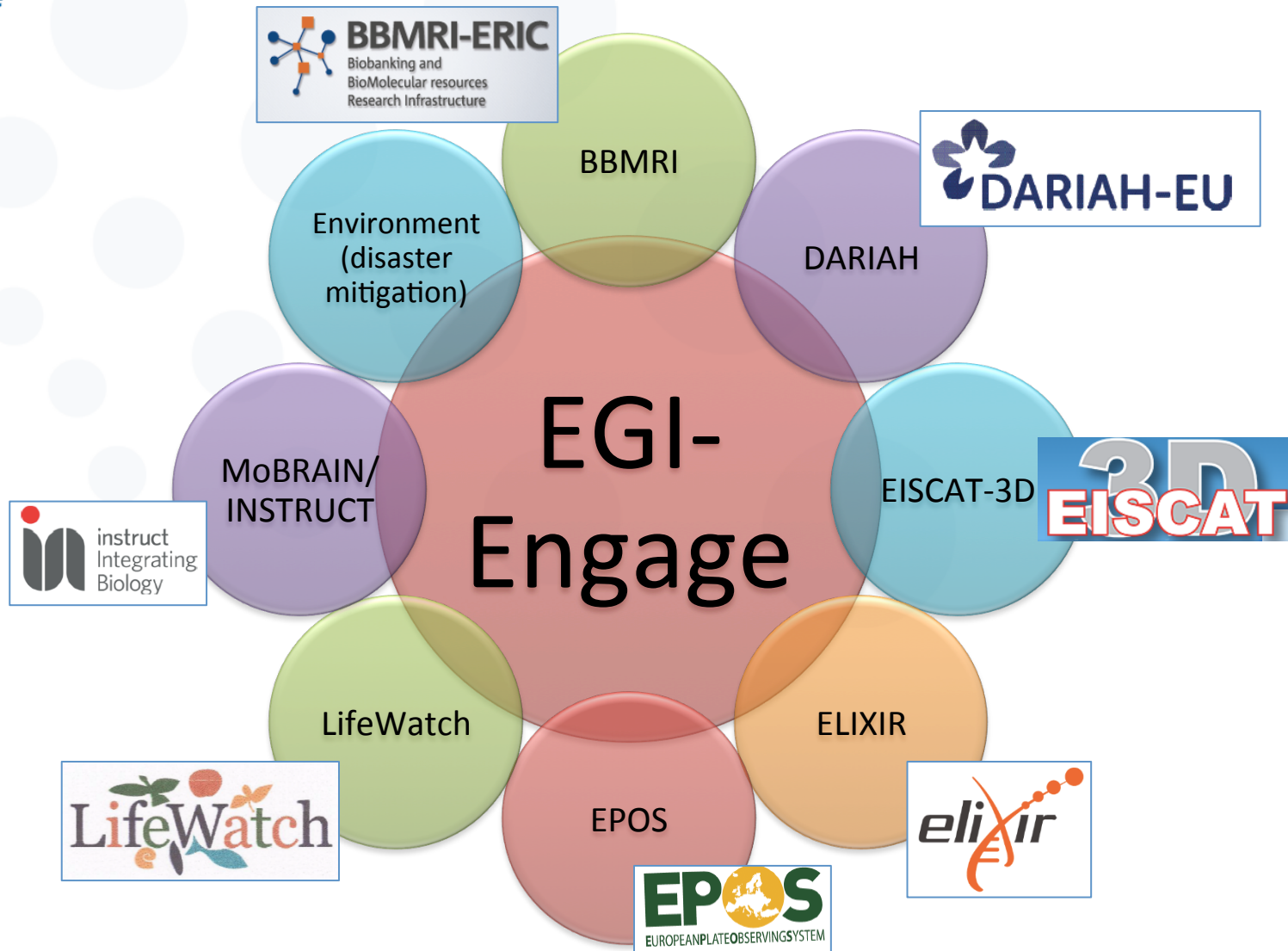- EDISON
- Technical support (ENVRI Plus, ELITRANS)

New governance to community engagement

The Distributed Competence Centre

Distributed Competence Centre (DCC)

- Promote reuse of solutions of common interest across research communities

- Evolve the EGI technical services with community requirements and provide a test environment with NGIs/EIROs → co-development

- Promote the integration of community services
  - Scientific applications
  - Joint training programme
  - Technical user support

EGI-Engage support to the DCC

Outcomes

Community clouds

Data products

Active Repositories

Training resources

GPU federations

Science gateways

Scalable applications

User support

Advanced AAI

Open Science

| MoBRAIN/ INSTRUCT | Environment (disaster mitigation) | BBMRI | DARIAH | EISCAT-3D | ELIXIR | LifeWatch | EPOS |

EGI community

| Federated Cloud | High Throughput Computing | Security, access control | Data management and federation | Gateway systems |

Actors – present and future

- Virtual Research Communities and supporting projects
- Research Infrastructures
- E-Infrastructures
- Centres of Excellence
- Technology providers
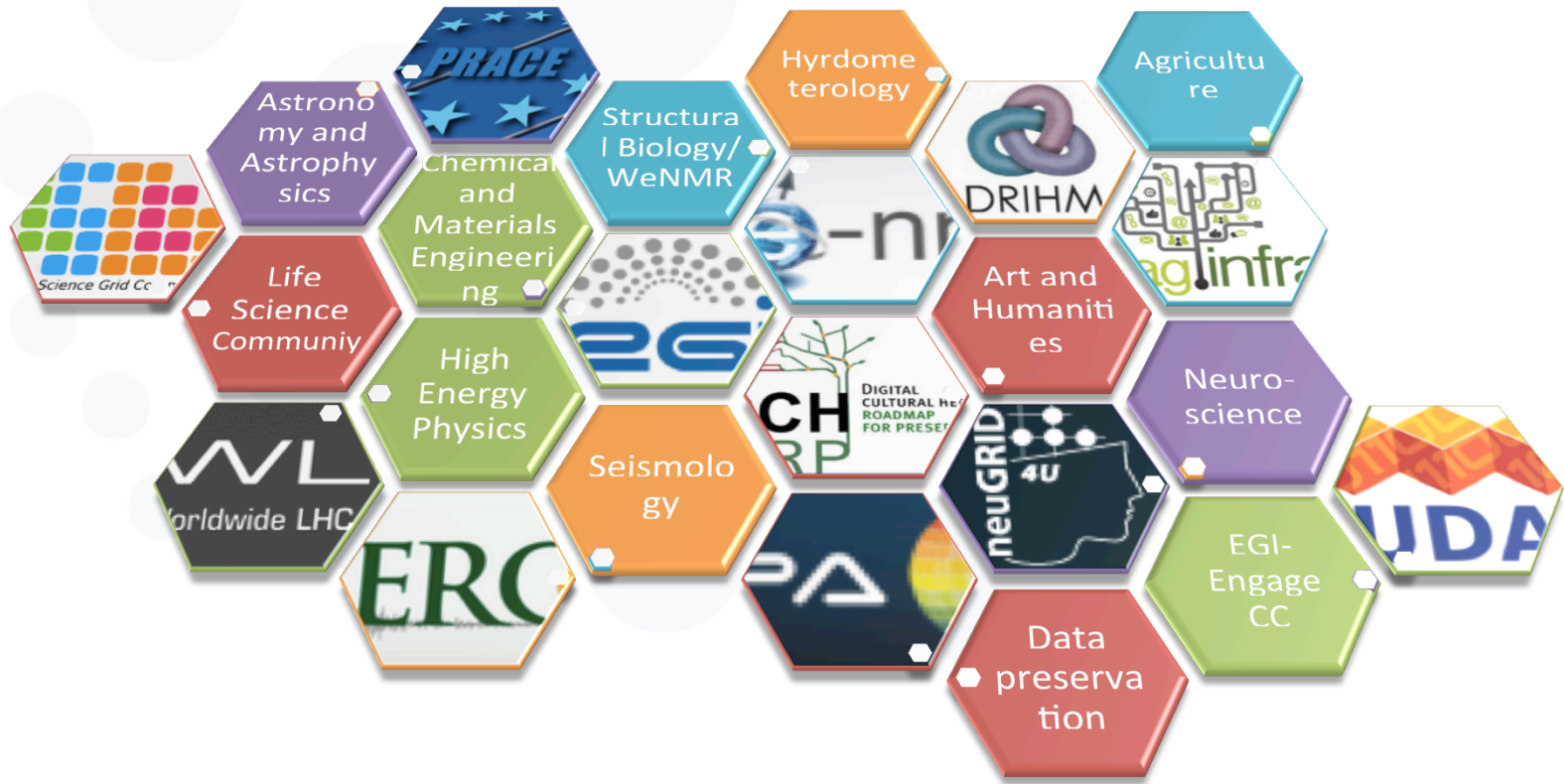- …

Federate Knowledge in Europe and beyond

Join the Competence Centre meetings, every day
17:00 – 18:00, OPEN!

New governance to community engagement

The Distributed Competence Centre

Better services for the long tail

Centrally provided services for reduced access barriers

Services for the long tail of science

- Move towards a "zero (technical) barrier" e-infrastructure
  - Services dedicated to individual users or very small collaborations:
    - No certificate, no VO, full EGI experience

- User facing features
  - Log in using their federated identity
  - Provide the additional information not available in the IdP
  - Discover (marketplace) and submit a request for resources

- EGI/NGIs facing features:
  - Assign UIDs to users of the long tail of science platform
  - Approve user request
  - Monitor usage of resources

The big shifts

New governance to community engagement

Better services for the long tail

New AAI

The Distributed Competence Centre

Centrally provided services for reduced barriers

Service proxy/ virtual IdP

Token translation

IdP for homeless users

Advancing AAI

- EGI users are directly/indirectly using x509 credentials to access the production services

- Objective: allow users to use their existing institutional credentials by
    - Replicating the current architecture to manage user communities in the other authentication technologies already used by the users
    - Integrating other federated identities into EGI services

- Testing and deployment of AAI services, and requirements analysis in close collaboration with the CCs and the other communities
    - Catch all IdP service (EGI sso), online CA, attribute authorities to manage users without X.509 ceritificate
    - Service proxy/Virtual IdP: technical service AND support to help communities to integrate easily their IdP with EGI. Integrating new IdP and attribute authorities in a one-step.

- Collaboration with AARC project

The big shifts

New governance to community engagement

The Distributed Competence Centre

Better services for the long tail
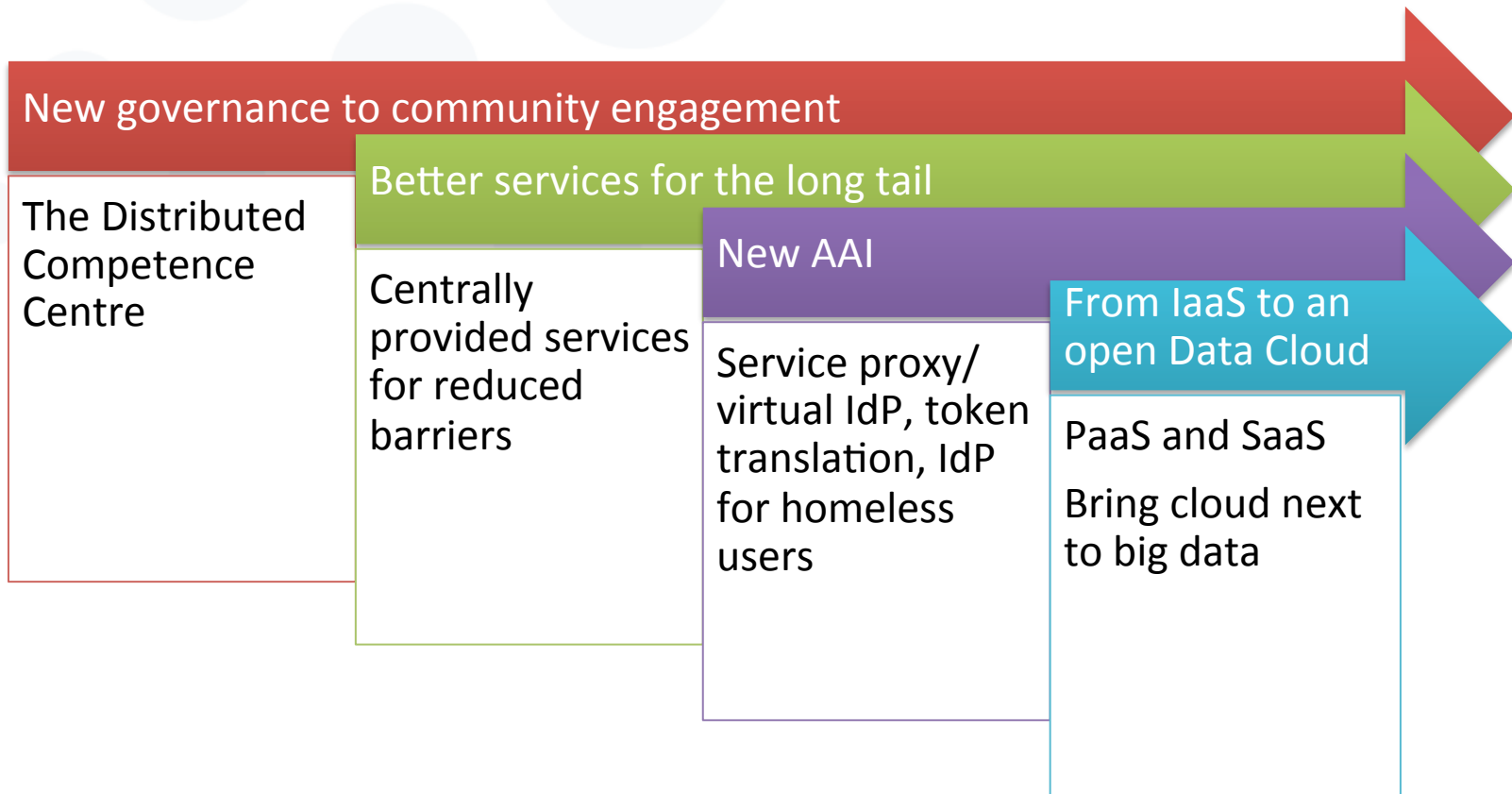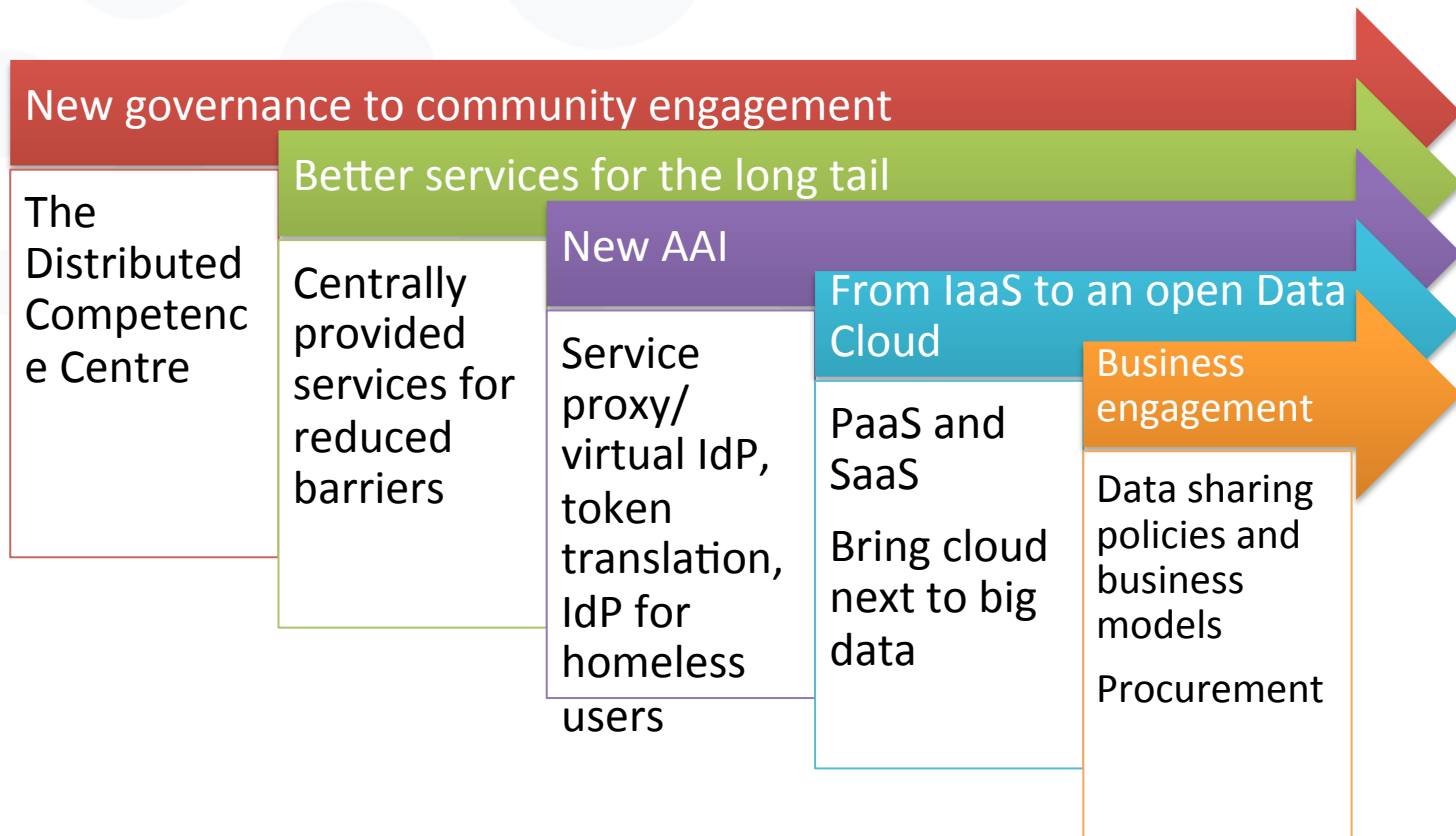
Centrally provided services for reduced barriers

New AAI

Service proxy/ virtual IdP, token translation, IdP for homeless users

From IaaS to an open Data Cloud

PaaS and SaaS

Bring cloud next to big data

Federated Cloud + Open Data: Open Data Cloud

- Objective: scalable access to open research data for discovery, access and use

- Remove policy and technical barriers
  - Bring cloud service next to distributed data repositories
  - Replicate open research data of research/commercial relevance
  - Discovery, accounting
  - Provide PaaS and SaaS and evolve the federation services
  - Virtual appliance library of community tools and data for
    - Repeatability of science, training and education (EDISON)

- Multiple stakeholders involved

The big shifts

New governance to community engagement

Better services for the long tail

New AAI

From IaaS to an open Data Cloud

Business engagement

| The Distributed Competence Centre | Centrally provided services for reduced barriers | Service proxy/virtual IdP, token translation, IdP for homeless users | PaaS and SaaS

Bring cloud next to big data | Data sharing policies and business models

Procurement |

Policy and business

- Pay-for-use and cross-border procurement
- Facilitate collaboration with SMEs (focus on consumer side) via a model to be adopted and adapted for a wider number of NGIs/Resource Centres
  - Use cases from agriculture, fishery and marine sciences, biodiversity, earth science
- Explore with SMEs opportunities and threats around the Open Data and co-develop business models for their exploitation
    - Market analysis and user requirements
    - Data Sharing Policies and Legal Aspects

Towards 2020

The big shifts

Achievements

EGI today
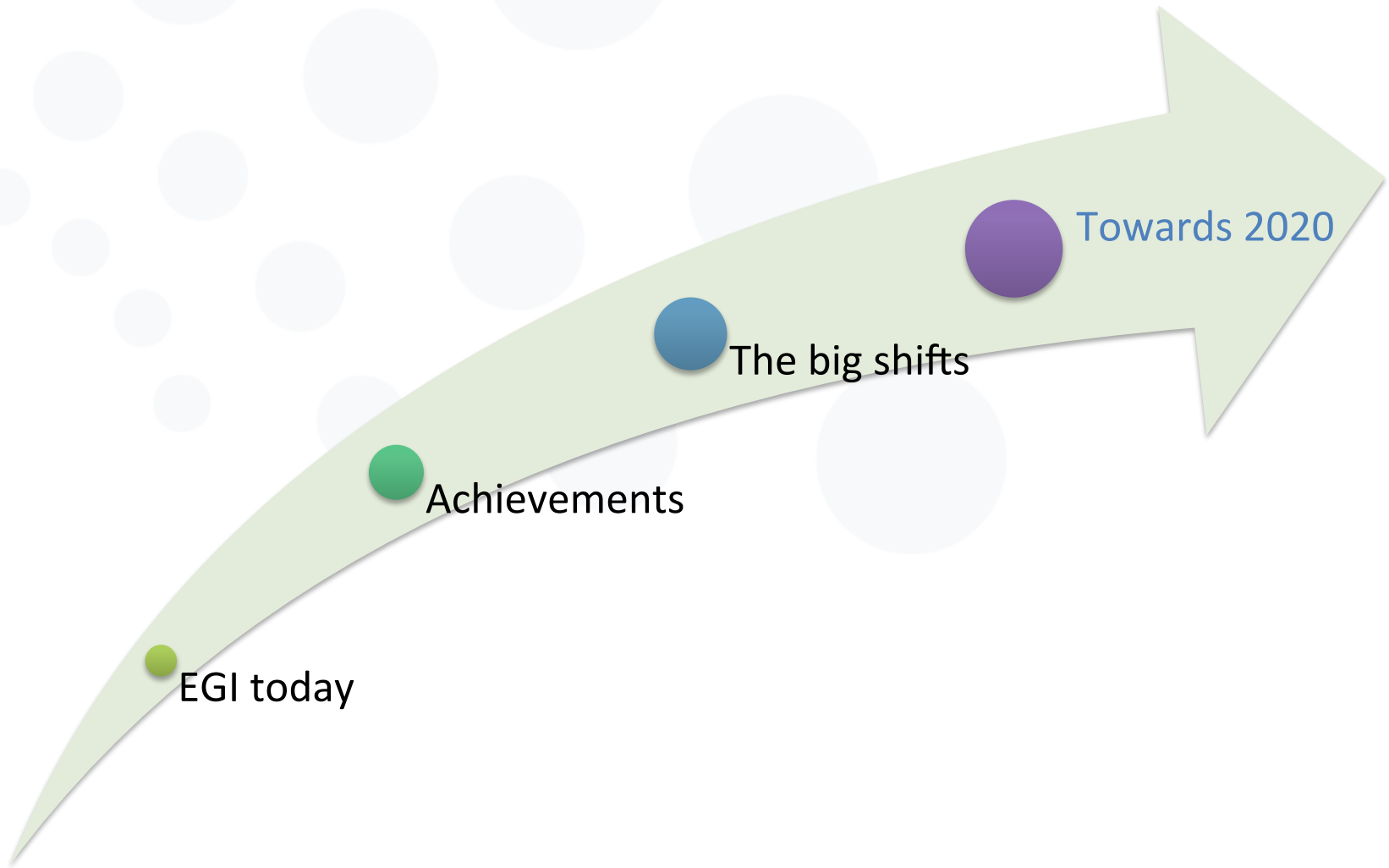
Digital ERA – State of play 2015

- Incomplete national roadmaps for Research and e-Infrastructures
  - E-Infrastructures and RIs should be components of the same research system

- e-Infrastructure Commons not fully achieved yet
  - Lack of e-Infrastructure capacity for multidisciplinary research and the long tail of science
  - Different access policies for user groups in each access
  - Incomplete technical interoperability, different access policies
  - The "Commons" governance principle not widely adopted
  - Non organized landscape of multiple service providers and research communities, lack of cross-border procurement/funding scheme that allows coordinated resource management across Europe (except for GEANT)

- Lack of one 'backbone' of European ICT capabilities

- Shared resources
  - Integrated, easy and fair access
- Engaged communities
  - Participating in the process
  - Culture of sharing
  - Collaborating in the management and stewardship
- Governance
  - Rules to access
  - Rules to resolve conflicts
  - Rules to balance quality vs. openness
- Financial support
  - For long-term availability

Research Data

Digital services and applications

Instruments

Knowledge & Expertise
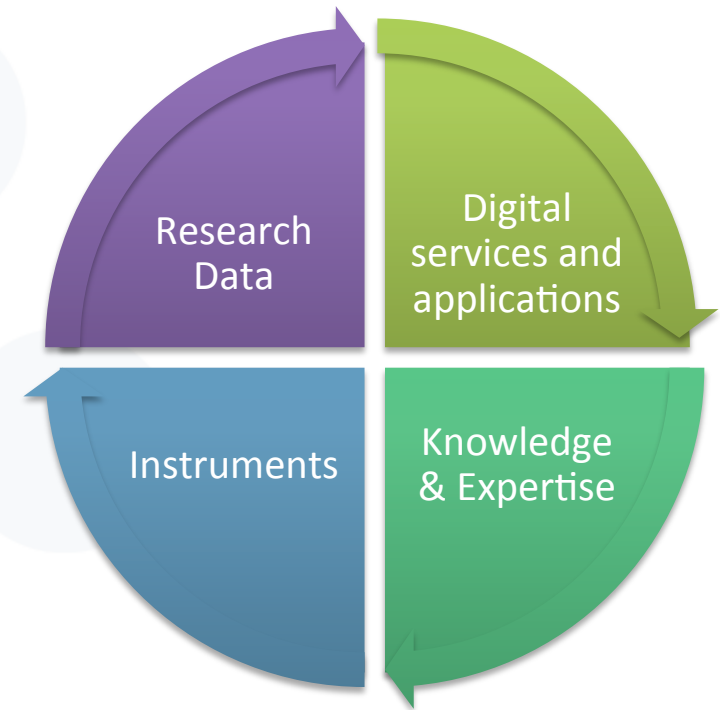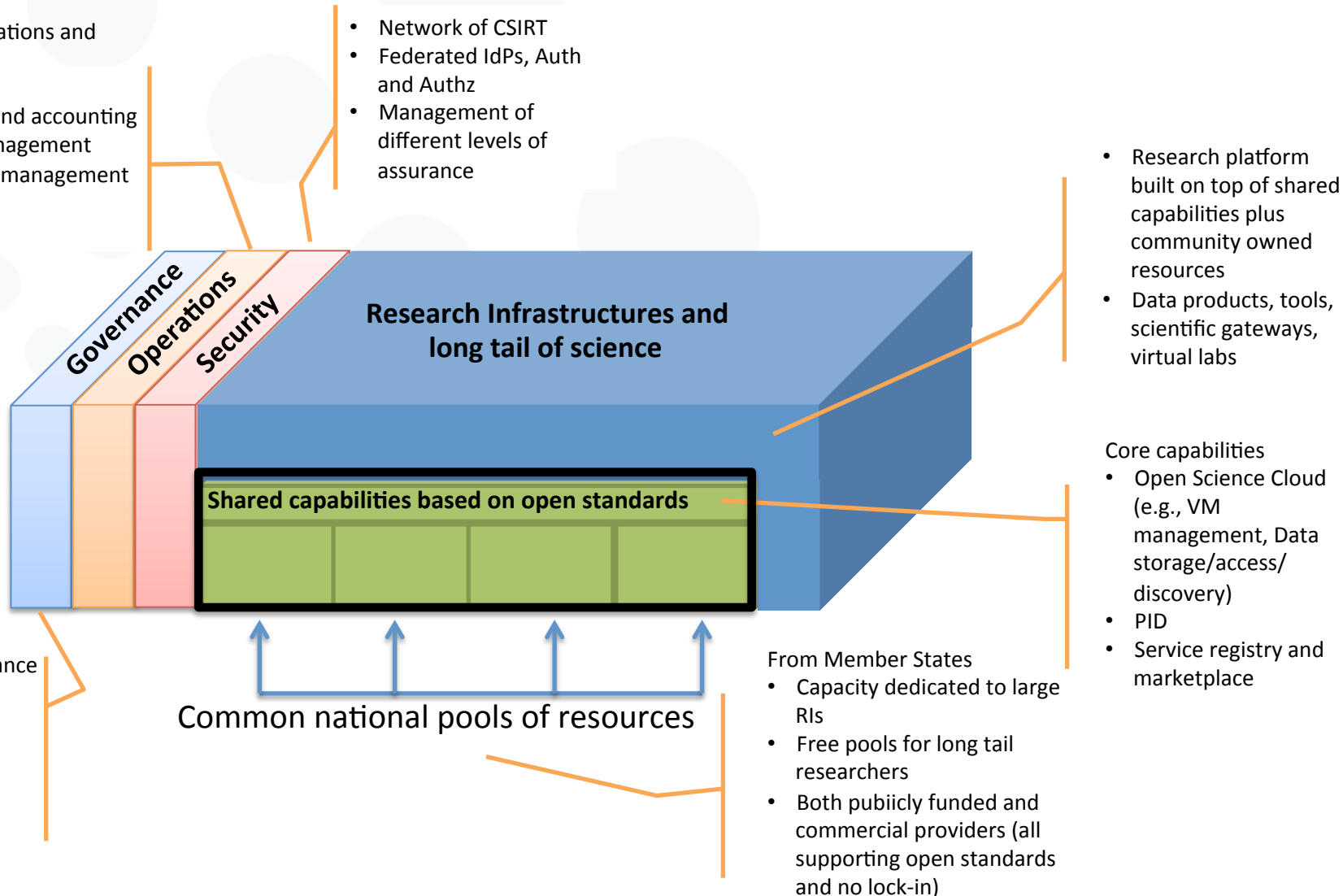
A common endeavor (EU perspective)

Developing an OSC:
Shared Open Science Infrastructure Backbone

EGI-Engage

Federated operations and support
- Service desk
- Monitoring and accounting
- Capacity management
- Service level management

- Network of CSIRT
- Federated IdPs, Auth and Authz
- Management of different levels of assurance

- Research platform built on top of shared capabilities plus community owned resources
- Data products, tools, scientific gateways, virtual labs

Governance

Operations

Security

Research Infrastructures and long tail of science

Shared capabilities based on open standards

Core capabilities
- Open Science Cloud (e.g., VM management, Data storage/access/discovery)
- PID
- Service registry and marketplace

Multi-level governance with community participation
- Local
- National
- European

Common national pools of resources

From Member States
- Capacity dedicated to large RIs
- Free pools for long tail researchers
- Both pubiicly funded and commercial providers (all supporting open standards and no lock-in)

How can EGI contribute?

Federate digital capabilities, resources and expertise

Operate services across the federated infrastructure

Co-create and integrate open and user-driven services and solutions

Be a trusted adviser on data and compute intensive science

Researchers from all disciplines
have easy, integrated and open access
to the advanced digital capabilities,
resources and expertise
needed to collaborate and to carry out
compute/data intensive science and
innovation

Create and deliver open solutions
for science and research
infrastructures
by federating digital capabilities,
resources and expertise
across communities and national
boundaries

Thank you for your attention.

Questions?