AuDaCE

AuDaCE

Projet d'infrastructure régionale pour le traitement des grands volumes de données scientifiques en Auvergne

V. Breton, CNRS-IN2P3
LPC Clermont-Ferrand, IdGC

Université Blaise Pascal Clermont-Ferrand UBP · irstea · cnrs · Université d'Auvergne · INRA Science & Impact · Inserm Institut national de la santé et de la recherche médicale · Clermont Université

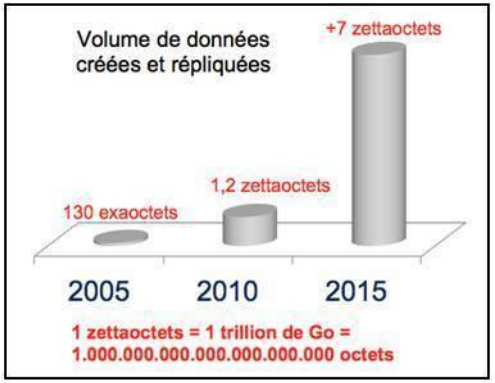
Introduction: projet AUDACE versus réseau AuDACES

- Coincidence fortuite
  - AuDACES = AUvergne, Développement d'Applications et Calcul en Environnement Scientifique
  - AUDACE = AUvergne Data Cloud académiquE

Le Big Data

Données numériques générées dans le monde:

- 2010: 1,2 zettaoctets (1 zettaoctet = 10^{21} octets)
- 2011: 1,8 zettaoctets
- 2012: 2,8 zettaoctets
- … 2020: 40 zettaoctets



Etude IDC Juin 2011 - The 2011 *Digital Universe* Study: Extracting Value from Chaos.

Données générées quotidiennement par:

- Twitter: 7 teraoctets (1 teraoctet = 10^{12} octets)
- Facebook: 10 teraoctets
- **Télescope LSST: 15-30 teraoctets par nuit**

- Le *Big Data* est généralement caractérisé par 3 ou 4 Vs:
  - Volume (données massives)
  - Variété (données complexes)
  - Vélocité (données à haut débit)
  - *Véracité (fiabilité des données)*

Quels sont les défis à relever?

- la collecte, le nettoyage, l'intégration et l'annotation structurale et/ou fonctionnelle des données,
- le stockage, l'indexation et l'interrogation des données,
- l'analyse et l'interprétation des données pour en dériver de la connaissance, et
- l'exploitation des connaissances dérivées pour une prise de décision rationnelle.

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

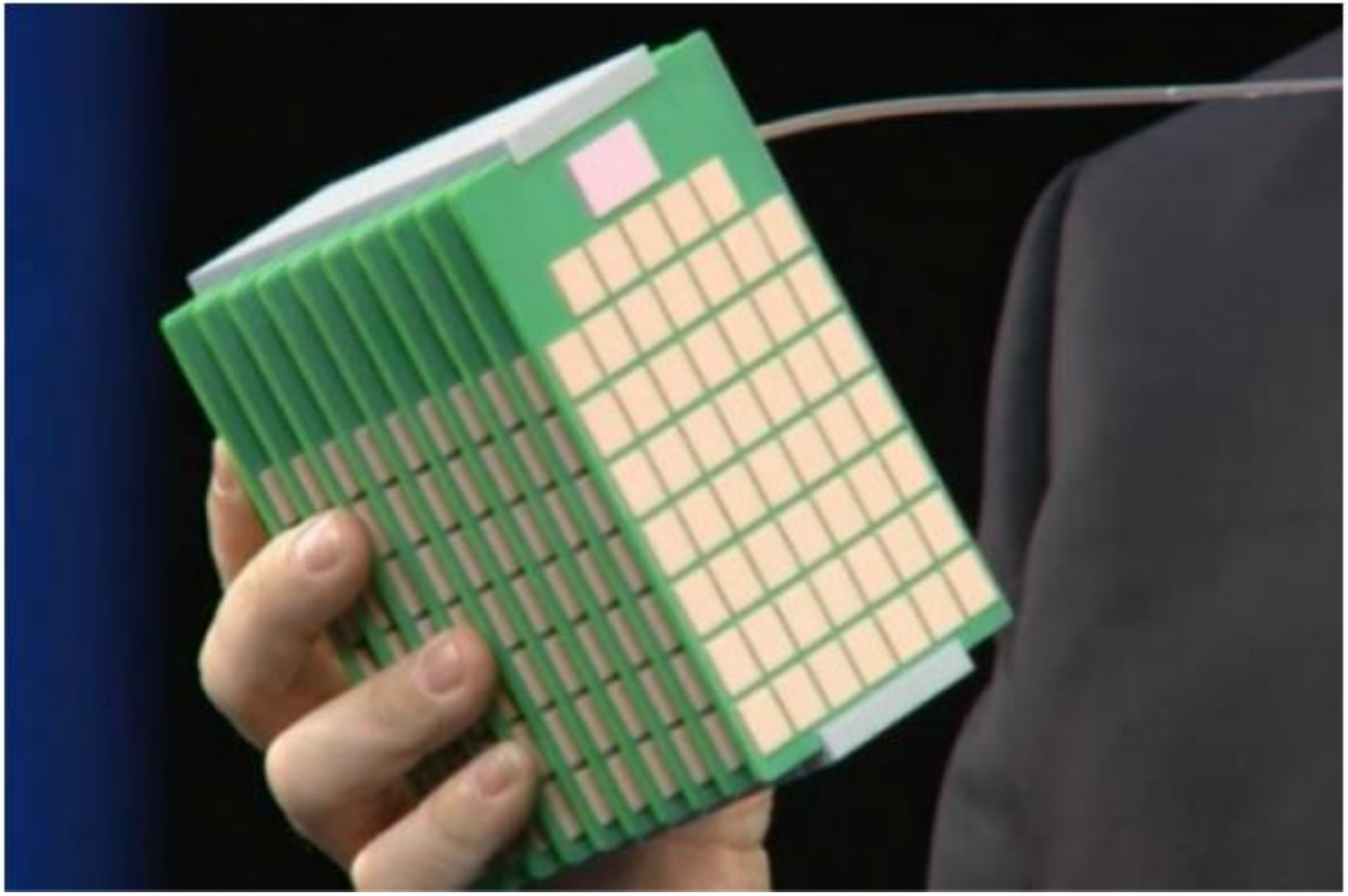
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

AUDACE

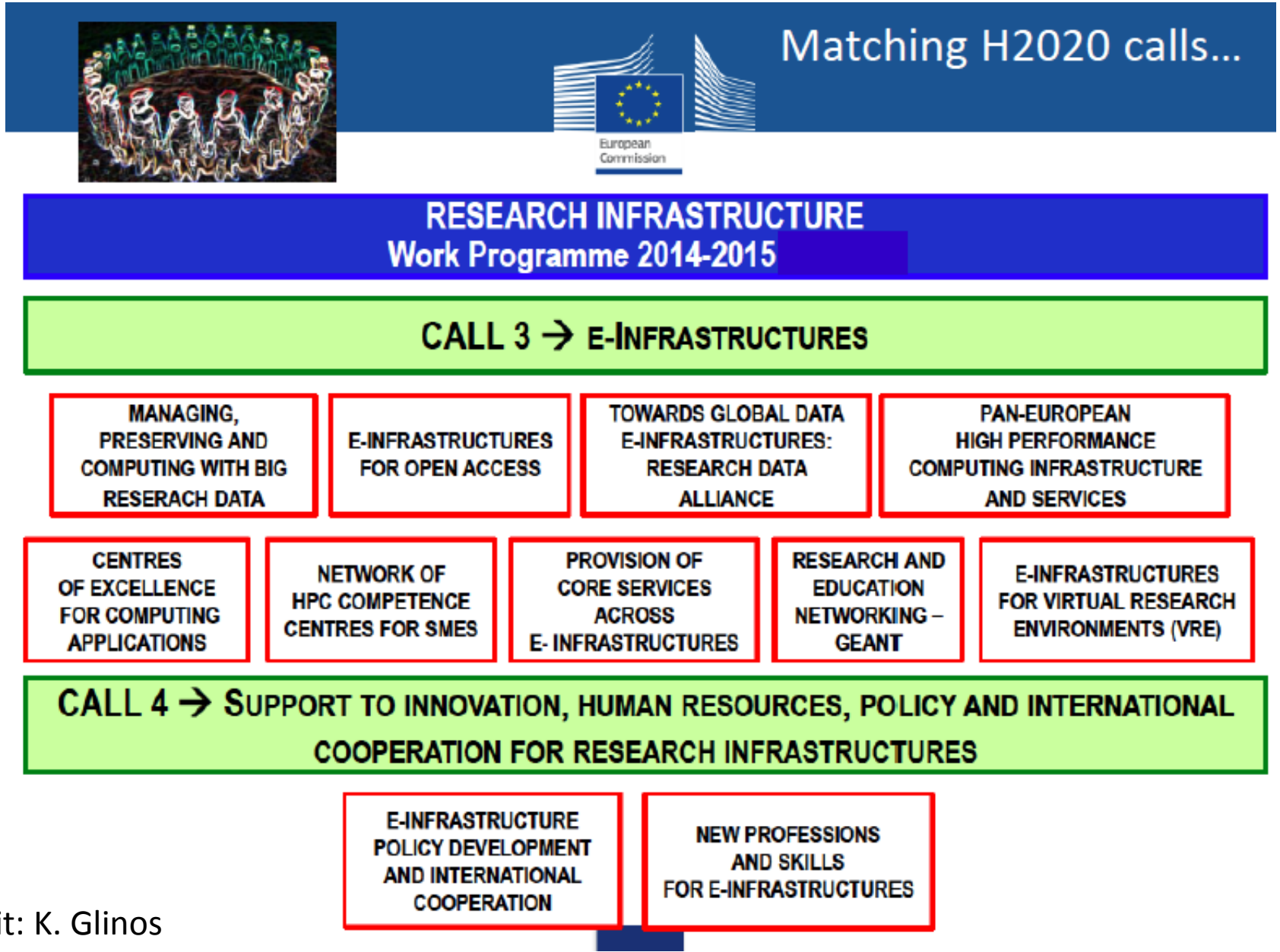
TECHNOLOGY

New Type Of Computer Capable Of Calculating 640TBs Of Data In One Billionth Of A Second, Could Revolutionize Computing

June 16, 2014 | by Justine Alford



Au cœur des appels à projets d'Horizon 2020



Credit: K. Glinos

Contexte international: une nouvelle organisation, la Research Data Alliance

Vision

Researchers around the world
sharing and using research data without barriers.

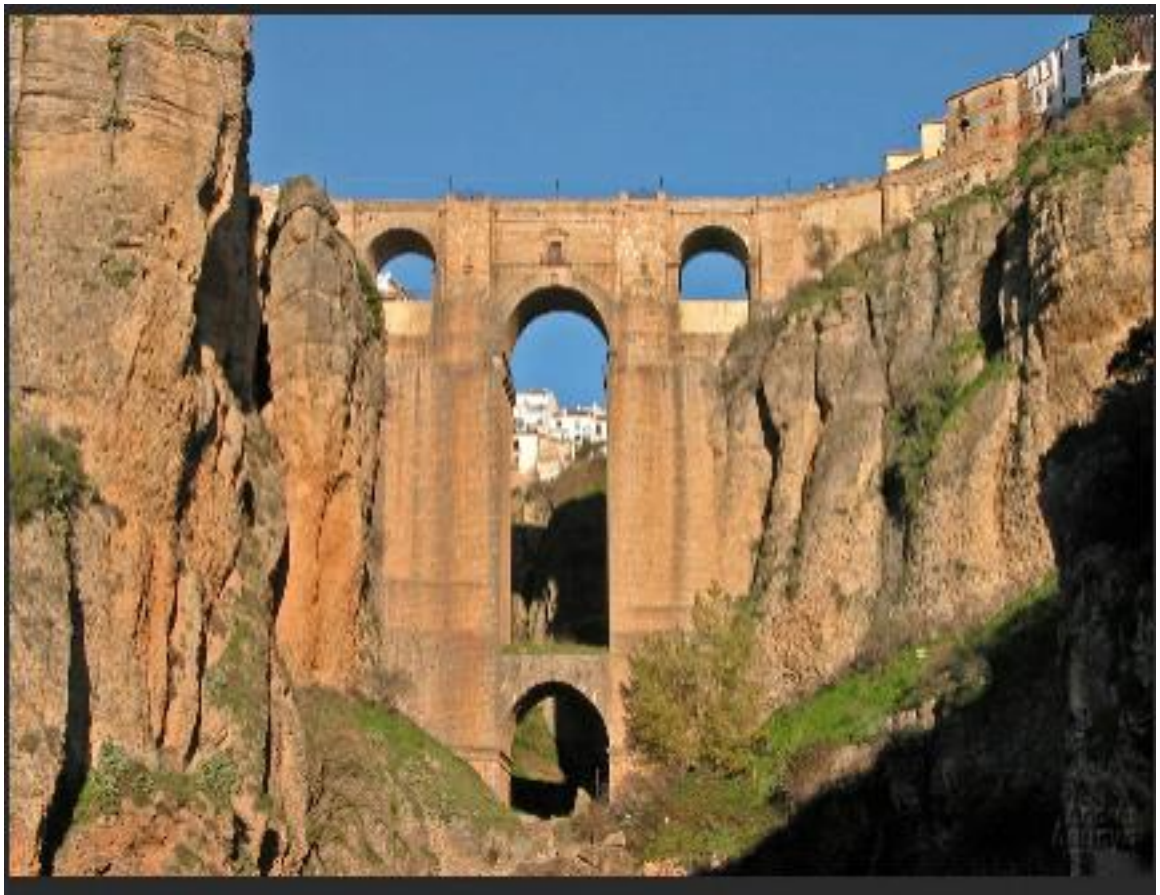
Purpose

... to accelerate international
data-driven innovation and discovery
by facilitating research data
***sharing** and **exchange**,*
***use** and **re-use**,*
***standards harmonization**, and **discoverability**.*
...through the development and adoption of
***infrastructure, policy, practice, standards**, and other deliverables.*

- Soutenue par la Commission Européenne, la National Science Foundation et l'Australian National Data Service
- Différent du Global Grid Forum

Objectifs de la Research Data Alliance

- Connecter les communautés d'utilisateurs
- Connecter les données



Research Data Alliance: construire des ponts

- Ponts vers le futur
  - Préservation des données
- Ponts vers les partenaires de la recherche
- Ponts à travers les disciplines
- Ponts vers l'intégration
  - Pour résoudre de nouveaux problèmes
- Ponts à travers les communautés

Plenary 4 - 22-24 septembre 2014 , Amsterdam,
L'évènement principal de la "Research Data Week"

Journée RDA-Europe du 20 juin 2014 au MENESR

AUDACE: construire les ponts… en Auvergne

Horizon 2020

AUDACE

Recherche en informatique

Centre Régional de Ressources Informatiques

Communautés scientifiques

AUVERGRID (CPER 2007-2013) – LIFEGRID (2006-2010)
INSTRUIRE (2005-2007)
ACI GRID (2002-2005)

Les objectifs du projet

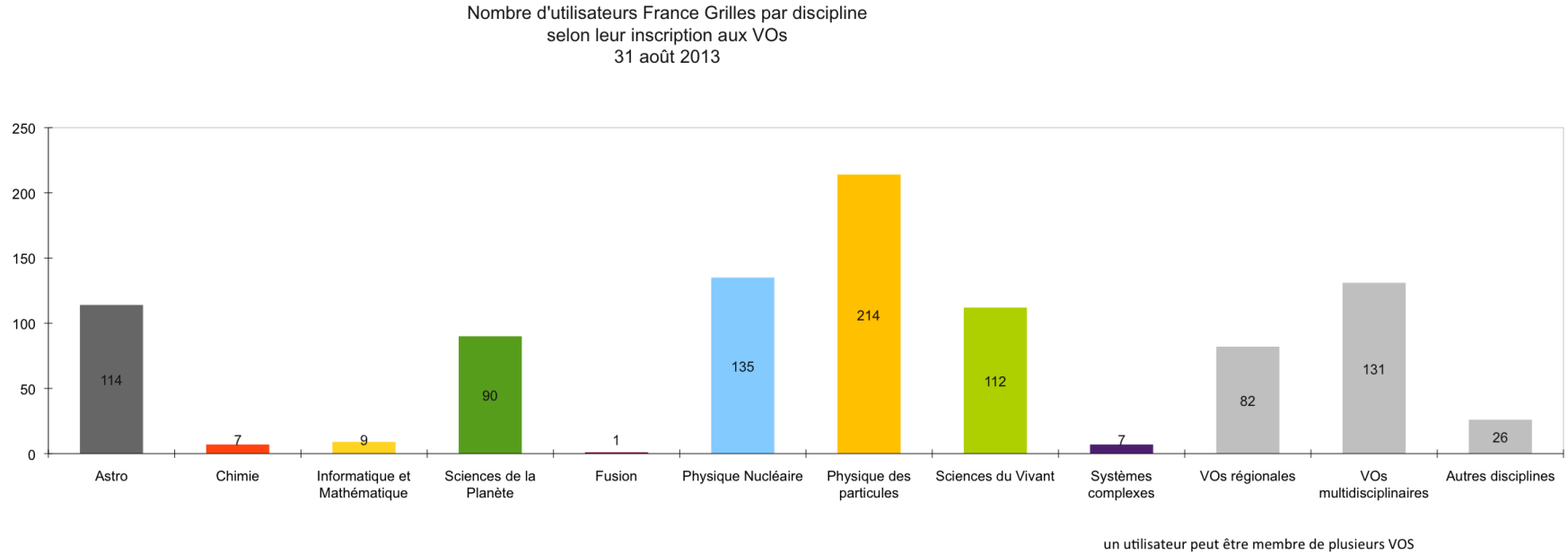
- Développer une recherche informatique originale sur le *Big Data*
  - Recherche générique
  - Recherche sur les données de grands instruments
  - Recherche sur les données liées à la politique de site
- Déployer une e-infrastructure pour les données scientifiques en Auvergne
  - Au service des communautés pour résoudre les défis scientifiques
  - Ouverte vers le monde socio-économique
  - Intégrée au niveau national et international

Une initiative intégrée dans France Grilles

Une initiative intégrée dans France Grilles

- France Grilles est un Groupement d'Intérêt Scientifique
  - créé en 2010 par 8 partenaires (CEA, CNRS,CPU, INRA, INRIA, INSERM, MESR, RENATER)…
  - pour animer et coordonner la stratégie nationale en matière de grilles et de clouds.
  - Son mandataire est l'Institut des Grilles et du Cloud du CNRS
- Sa vision:
  - Construire et opérer une infrastructure informatique distribuée ouverte à toutes les sciences et aux pays en développement qui constitue un espace ouvert de collaboration au sein et entre les disciplines

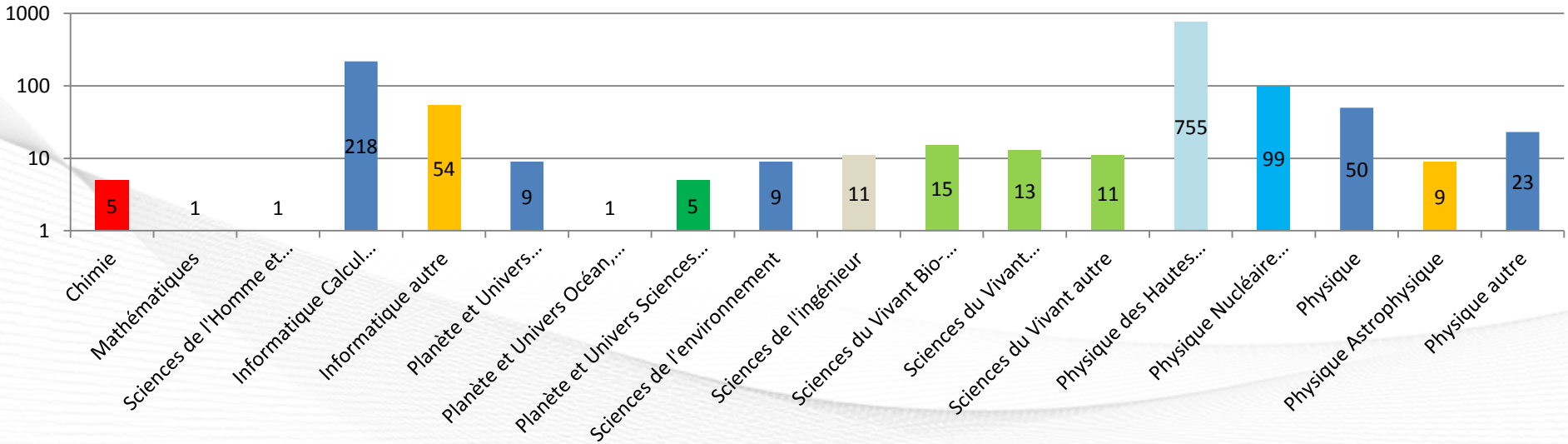
Qui sont les utilisateurs des ressources et services ?



-Plus de 750 utilisateurs
-une vraie pluridisciplinarité
-un flux régulier de nouveaux utilisateurs
-une évolution des usages avec l'offre de services

Importante production scientifique

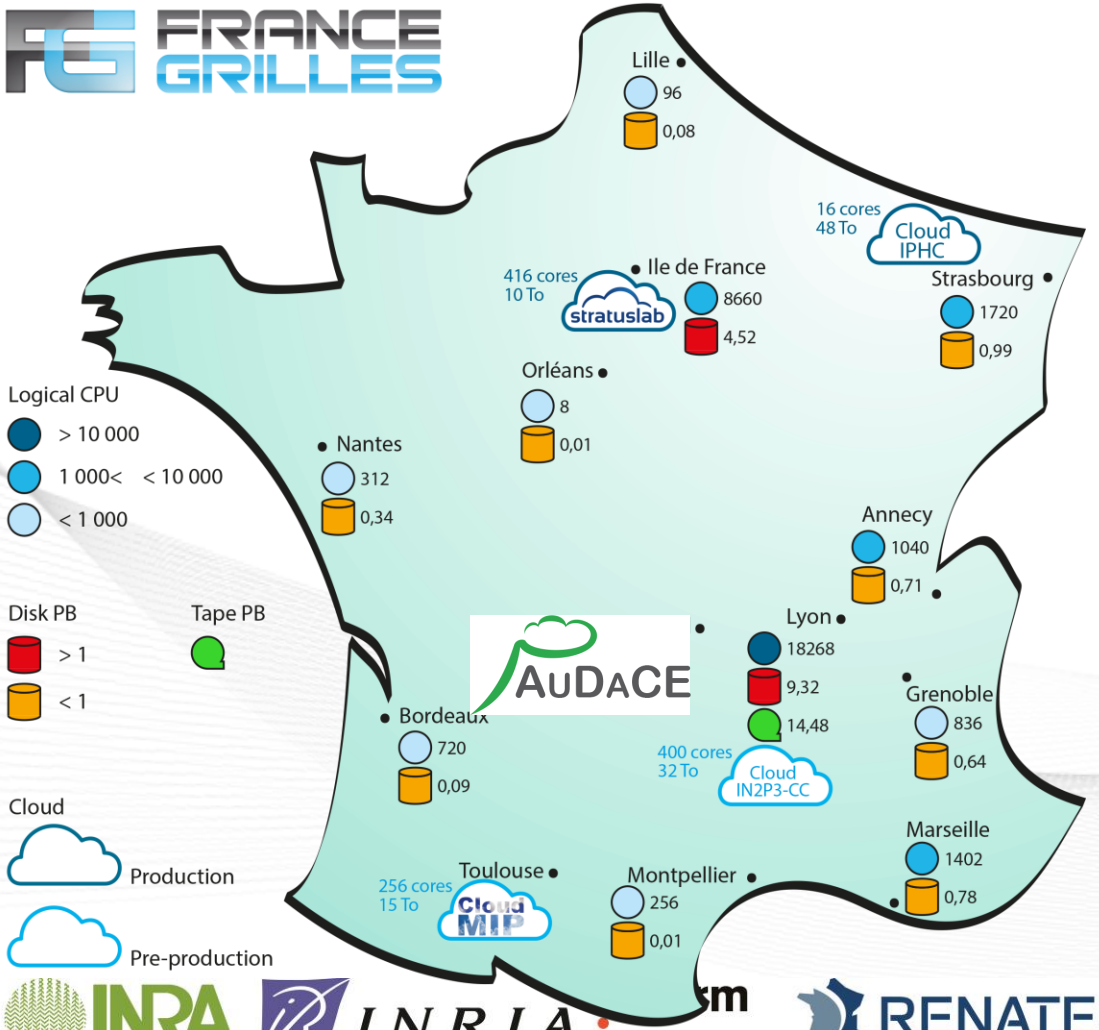
Plus de 1600 référencements dans HAL
1150 de Avril 2010 à Avril 2014



Et un prix Nobel !

Crédit: Geneviève Romier

AUDACE fournisseur de ressource dans France Grilles



Colonne vertébrale de France Grilles: LCG-France

Coordination technique: CC-IN2P3

Organisation du projet

Recherche générique *Big Data*

Sciences de la vie et de la santé

Microbiome

Données géoréférencées

Astrophysique (LSST)

Axe I – EPICURE Sciences biomédicales

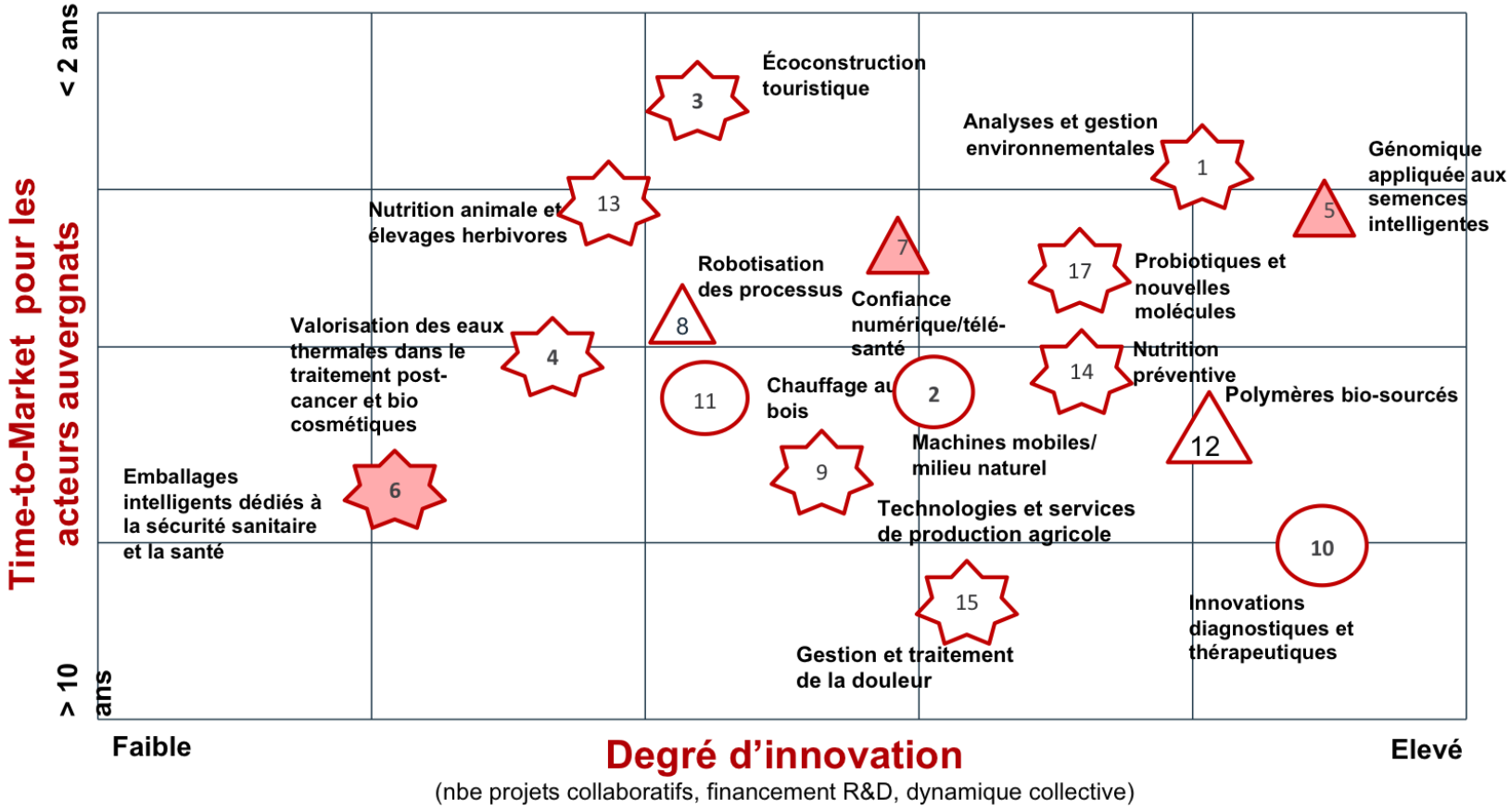
Axe II – SYMBIOSE Sciences de l'environnement

Axe IV – ATTRIHUM Sciences sociales

Axe III MMaSyF Sciences pour l'ingénieur

CRII – mésocentre régional

Au cœur de la problématique sur la traçabilité et la sûreté physique et numérique du vivant, des produits et des données de la Stratégie de Spécialisation Intelligente



Résultats attendus

- Activité scientifique reconnue autour de la gestion des grands volumes de données scientifiques
- Accroissement de la compétitivité de la recherche en Auvergne, notamment dans les quatre axes identifiés de la politique de site
- Fédération des acteurs académiques en région autour de la construction d'un catalogue d'offres de services et du partage de références autour de la problématique *Big Data*
- Montée en puissance du CRRI dans son rôle de mésocentre et de centre de gravité des services informatiques pour les acteurs académiques régionaux
- Insertion du mésocentre régional dans la pyramide national du calcul intensif
- Participation significative à des projets/e-infrastructures nationaux ou internationaux

Domaines d'excellence

- Recherche générique sur le *Big Data*
- Recherche sur les données du téléscope LSST
  - Coordination du projet PetaSky (CNRS)
- Problématiques *Big Data* spécifiques
  - Données géoréférencées
  - Ecosystèmes microbiens (Méta-omique à haut débit)
  - E-santé
- Participation d'acteurs « locaux » à des projets nationaux et européens d'e-infrastructure
  - France-Grilles et EGI
  - LifeWatch

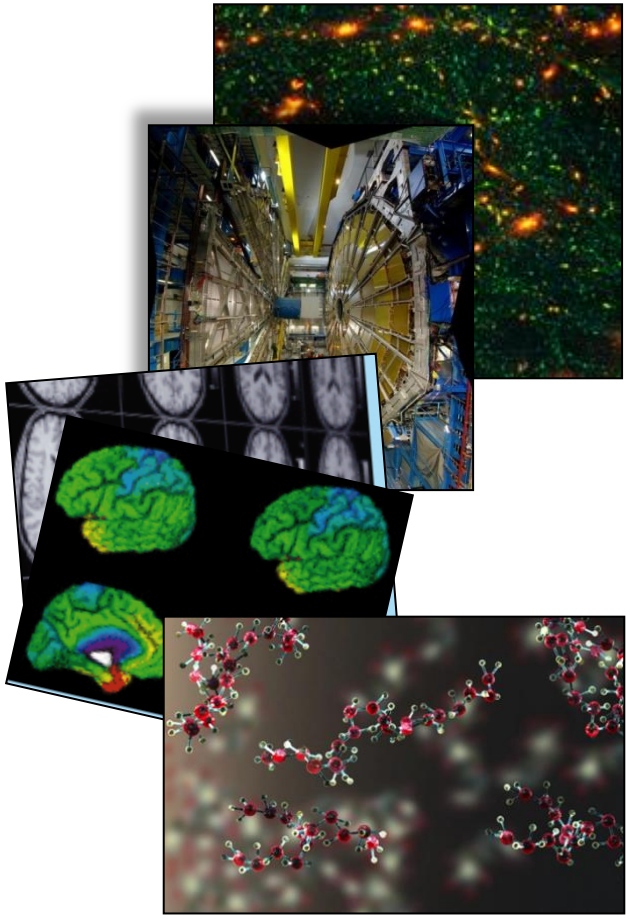
Intégration internationale

EGI Mission

MISSION: To support international researcher collaborations from all disciplines with the reliable and innovative ICT services they need to accelerate excellent science

- Natural and physical sciences
- Medical and health sciences
- Engineering and technology
- …

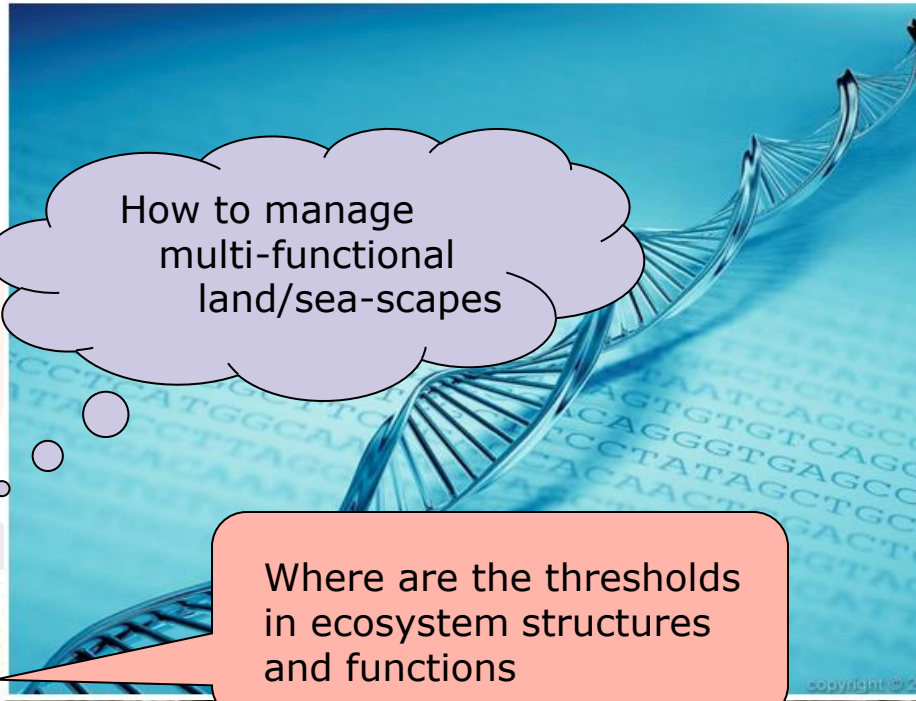
EC **EGI-InSPIRE** project (2010-2014)

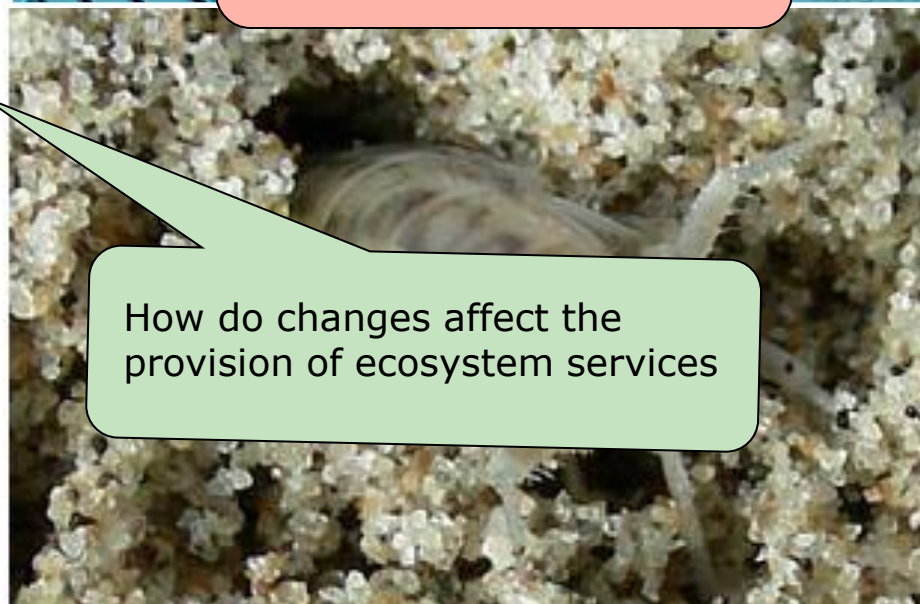


http://www.egi.eu/case-studies/

LifeWatch

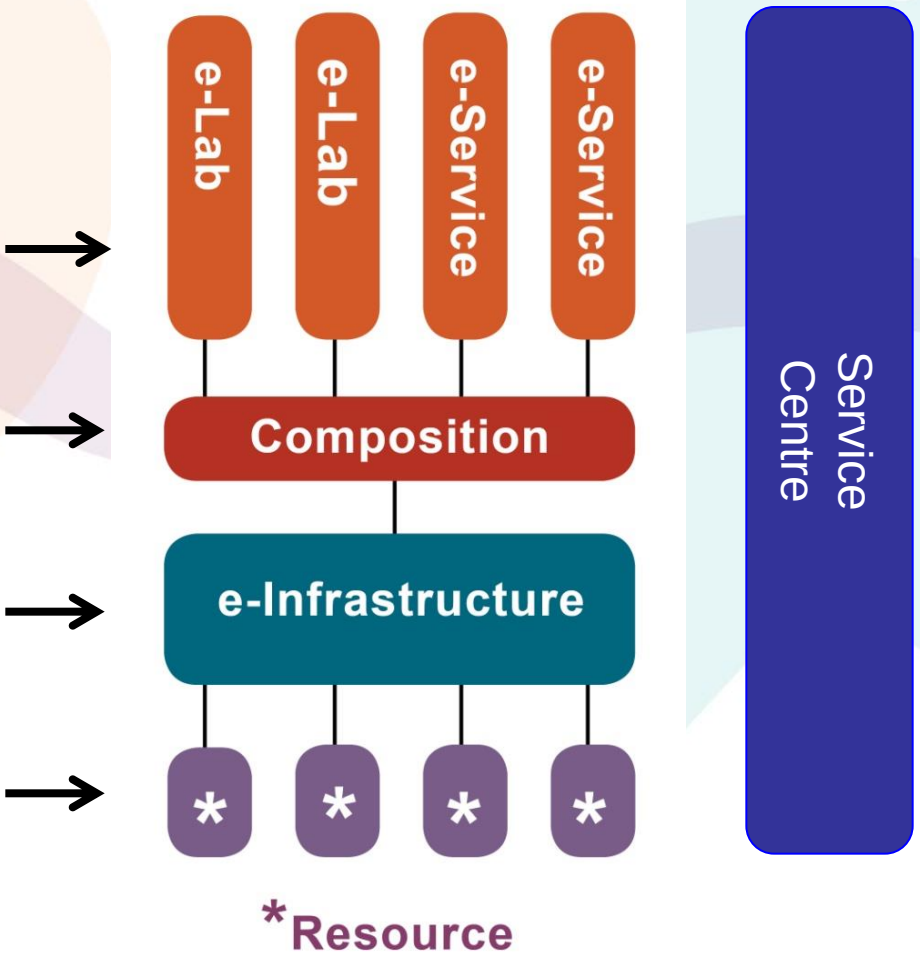

Which actions to ensure long-term sustainability
Can we adapt to environmental change
?
What are the impacts of changes in climate, pollution and land/sea-use on biodiversity


How to manage multi-functional land/sea-scapes
Where are the thresholds in ecosystem structures and functions


How do changes affect the provision of ecosystem services

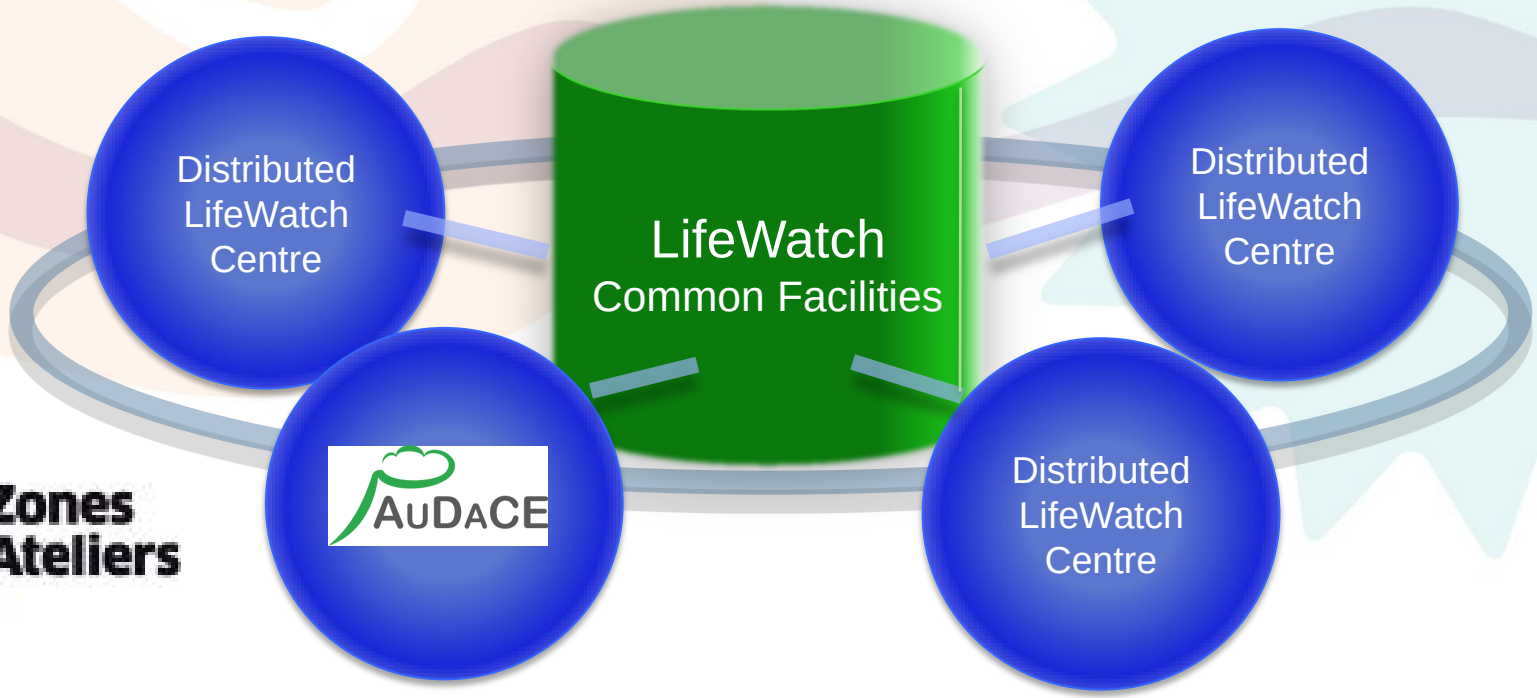
LifeWatch architecture

- Virtual laboratories for scientific cooperation
- Select the data, software, computing power
- Integrate resources
- Linking to resources (databases, sensors, software, computing power)



Contribution de AUDACE à LifeWatch

LifeWatch is cooperating with "distributed" LifeWatch Centres in cooperating countries, operating parts of the facilities and services.



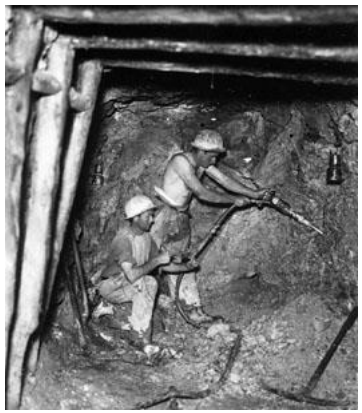
All together these constitute the "LifeWatch Research Infrastructure"

Auvergne at the heart of Uranium production in France

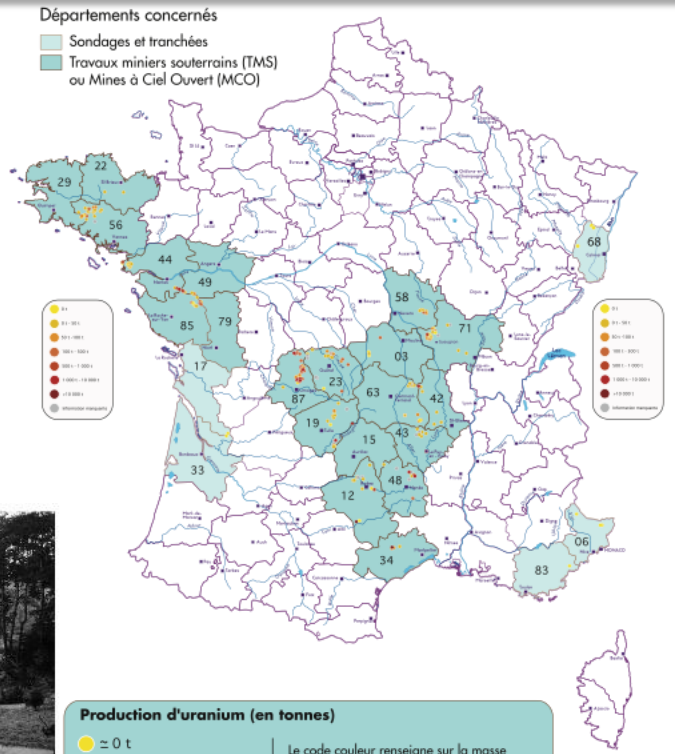
1949: first attempt to extract uranium ore in France in Lachaux (Auvergne)

In 50 years:

- 53 Million tons extracted in France till 2001
- 76000 tons of uranium ore produced in > 200 mines



Map of uranium mines in metropolitan France



LTER FRANCE

ZATU, a Long Term Ecological Research dedicated to life under natural ionizing radiation

Natural radioactivity

Storage sites of uranium ore extraction residues

- Society in uranium rich territories
  - Social impact of uranium extraction
  - Preserving the long term memory
- Characterization, behavior and transfer of radionucleids
  - long term future of radionucleids in storage sites
- Impact of radiation on living systems
  - Multigenerational effects of chronic exposure to radiation

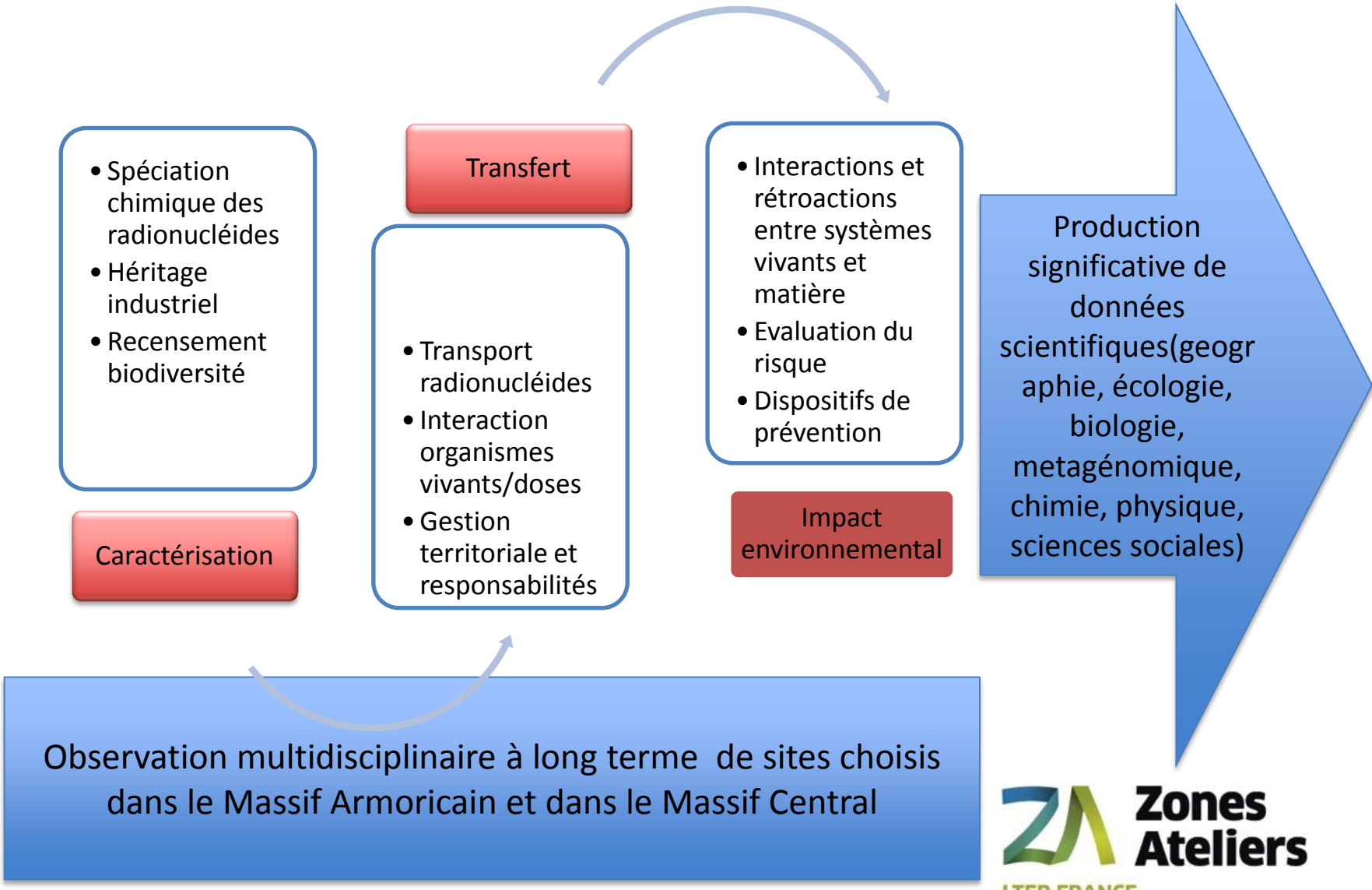
Impact of chronic exposure to low dose ionizing radiation on living organisms

- From the Chernobyl environment, a coherent picture of predictable radiation-induced effects for low-dose-rate exposures has not emerged
  - Contradictory experimental evidences from Chernobyl exclusion zone
- Need to collect more data from Chernobyl exclusion zone but also from other ecosystems under chronic low dose exposure

Photographs of abnormalities in barn swallows. (a) Normal phenotype. (b–d) Partially albinistic plumage. (e) and (f) Deformed beak. (g) Deformed air sacks. (h) and (i) Bent tail feathers.

Proasellus cavaticus

Stratégie d'étude des écosystèmes sous irradiation chronique



Zones Ateliers
LTER FRANCE

Phasages du projet

- Phase I (2015-2017)
  - Recherches génériques sur le *Big Data*
  - Recherches *Big Data* en astrophysique, sur le microbiome, les données géoréférencées et biomédicales
  - Déploiement d'une e-infrastructure de services pour le *Big Data* centrée au CRRI
- Phase II (2018-2020)
  - Poursuite des recherches *Big Data* génériques
  - Extension à de nouvelles thématiques de recherche sur des données thématiques
    - Sciences humaines, Sciences du vivant, Sciences de l'Ingénieur
  - Développement d'un écosystème de recherche et développement sur l'e-infrastructure pour le *Big Data*

Leviers à actionner

- Faire du CRRI un véritable mésocentre
  - Une offre de services pour le calcul scientifique
  - Une masse critique d'expertise
  - Une gouvernance qui implique les communautés d'utilisateurs
- Des choix technologiques en cohérence avec
  - les instances nationales (GENCI, GIS Grid5000 et France Grilles)
  - les communautés scientifiques (bioinformatique, physique des particules,…)
- Une participation à des projets internationaux (LSST) et des Infrastructures européennes (lifeWatch, ELIXIR)
- Une offre de formation initiale et permanente

Budget

- Demande: 7,7 Millions d'Euros
  - Equilibre recherché: 2/3 Ressources Humaines, 1/3 équipement informatique
- Besoins en matériel structurant au CRRI
  - Construction d'une offre de service
- Besoins en ressources humaines
  - Thésards et post-doctorants pour accompagner et alimenter les recherches sur le *Big Data*
  - Ingénieurs pour construire et développer l'écosystème d'une e–infrastructure en Auvergne

AuDaCE

Conclusion

- Dynamique autour de la problématique Big Data
  - En Europe
  - En France
  - En Auvergne
- Le projet AUDACE (CPER 2015-2020) a 2 objectifs principaux
  - Offrir des ressources et des services pour le traitement de grands volumes de données
  - créer des ponts au niveau local entre les communautés scientifiques et le monde socioéconomique
  - … au coeur des dynamiques nationales et internationales