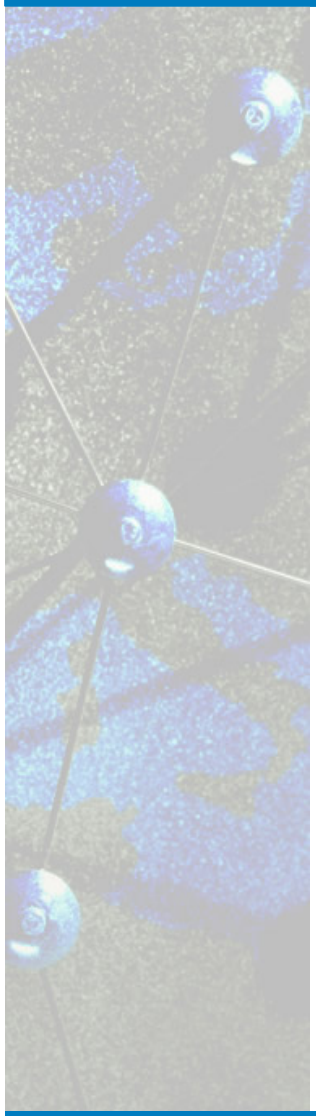


EGI-InSPIRE Requirements & Long-term Needs for Heavy User Communities

Maria Girone, CERN-IT &
EGI-InSPIRE

e-infrastructure

Introduction

- This talk will describe on-going multi-disciplinary work aimed at identifying and providing common, sustainable solutions where possible

- Building on concrete results to date it will examine future prospects particularly in data archival and preservation areas

The Disciplines

- We will talk briefly about the needs from Astronomy & Astrophysics, Life & Earth Sciences as well as High Energy Physics

- These disciplines have widely varying computing models but nonetheless use a common production infrastructure – albeit with significant customization and / or higher level components

- The goal today: identify potential solutions that can be applied even more widely in the MSST domain

Motivational Factors

1. EGI-InSPIRE: a 48-month project funded by the European Union with explicit support for a number of “Heavy User Communities”
   – But for 36 months only…

2. Framework Programme 8 – the next round of projects / funding – with a significant focus on “The Digital Agenda for Europe”

- We all know that we are living in an increasing digital world – the goal is to go way beyond “early adopters” to all aspects of science and society

Project Overview

- EGI-InSPIRE: a multi-disciplinary EU grid project involving:
  - High Energy Physics (HEP);
  - Life Science (LS);
  - Earth Science (ES);
  - Astronomy & Astro(particle)physics (A&A);
  - Computational Chemistry;
  - Fusion;
  - Plus tools and services used by multiple communities

- Provide support to current structured international research communities using:
  - Commodity or High Performance and Throughput clusters
  - Commodity and / or dedicated high-performance network links
  - Disk or tape storage
  - Data Archives or Digital Libraries

- Continue towards a sustainable production infrastructure

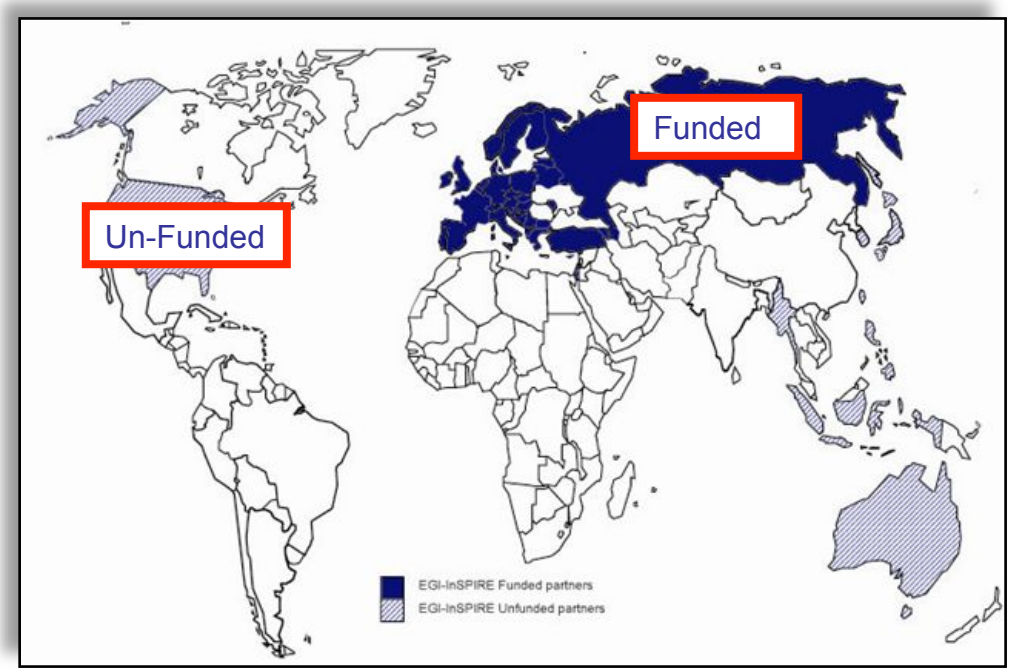
EGI-InSPIRE Project

Integrated Sustainable Pan-European Infrastructure for Researchers in Europe

A 4 year project with €25M EC contribution

- Project cost €72M
- Total Effort ~€330M
- Effort: 9261PMs

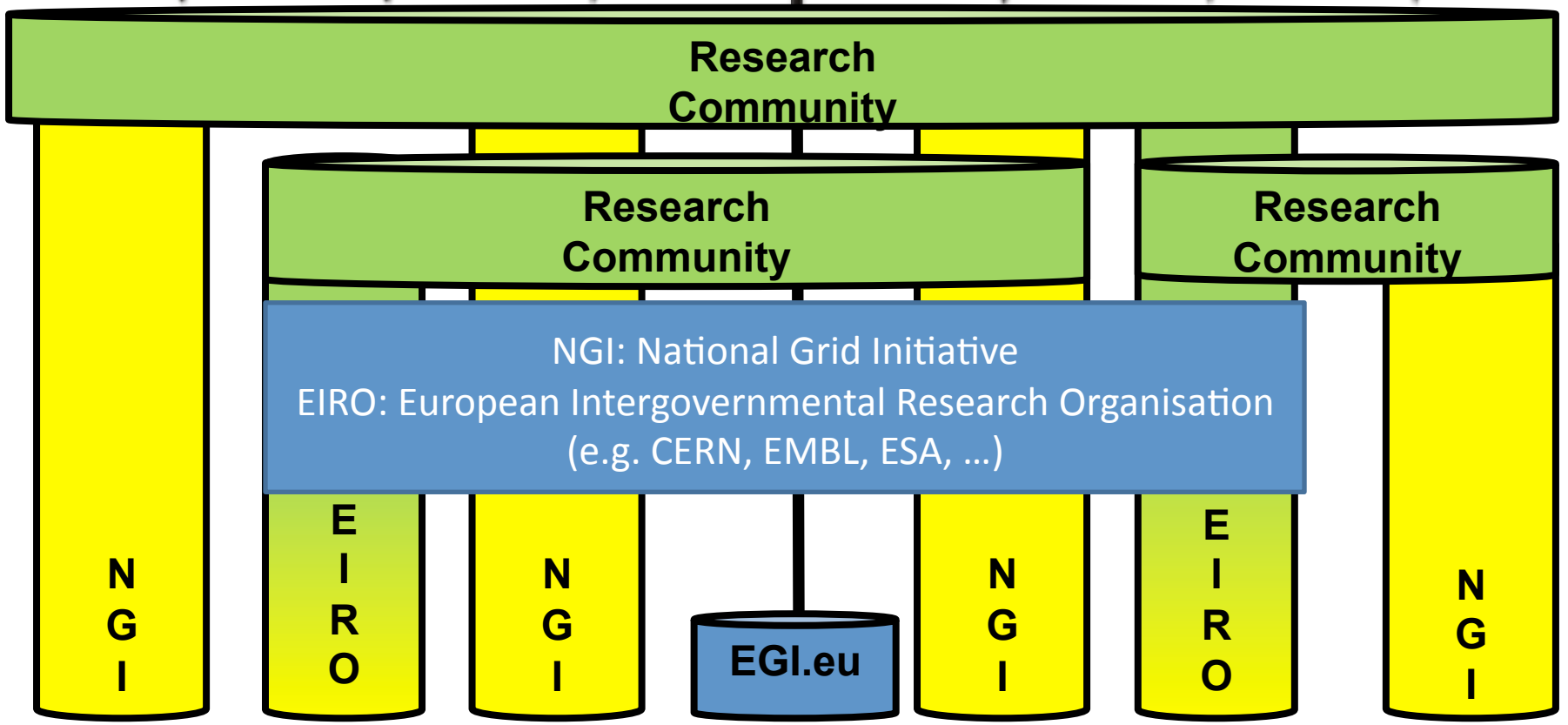
Project Partners (50)
EGI.eu, 38 NGIs, 2 EIROs
Asia Pacific (9 partners)



HUC Activities

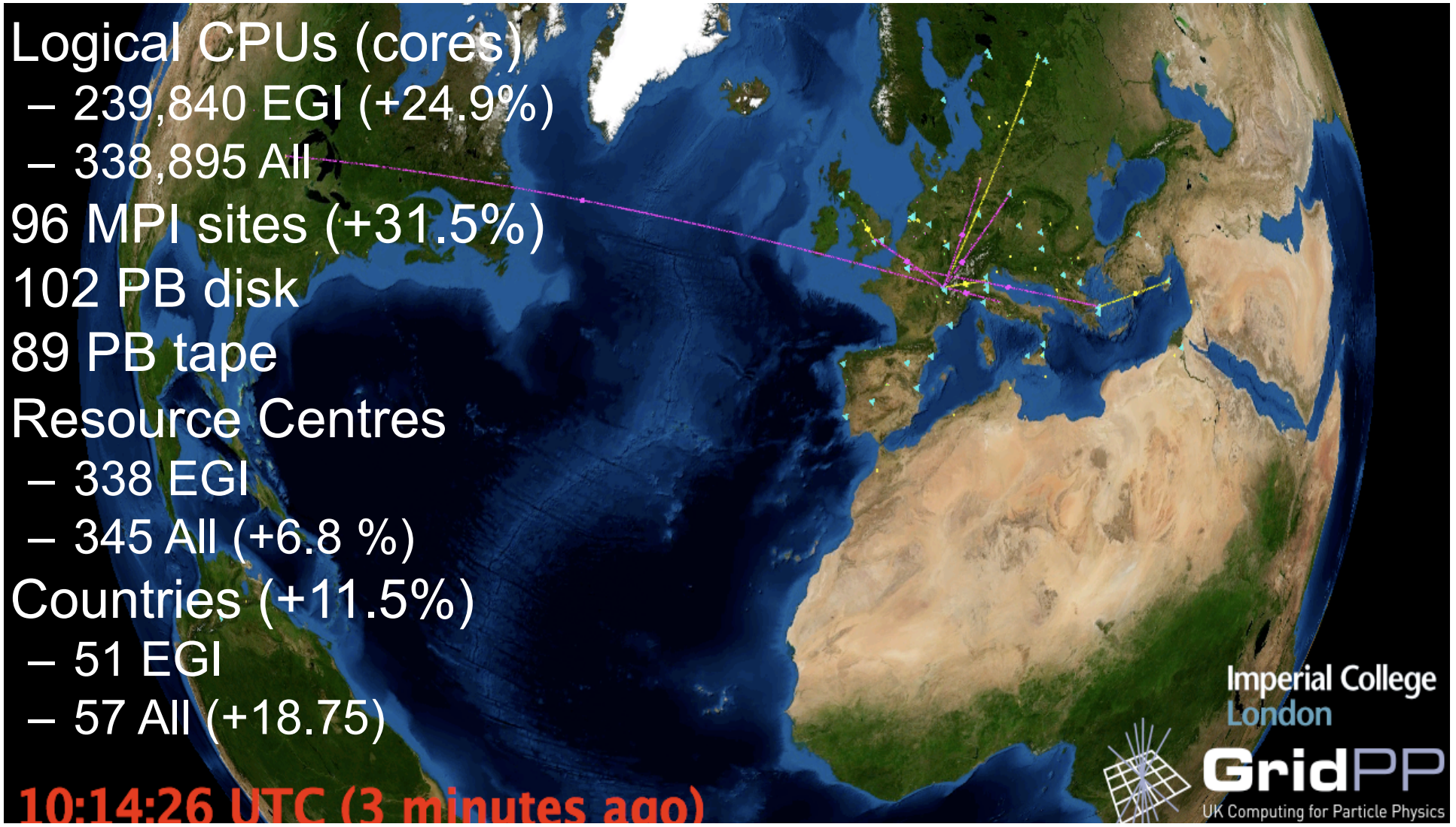
Community	Description, capabilities and services offered
All HUCs	This task provides support for **grid** tools and services that are used by more than one community, including Dashboards, applications such as Ganga, Services such as HYDRA and GRelC, Workflows and Schedulers (SOMA2, Kepler, Taverna) and MPI.
High Energy Physics	The High Energy Physics (HEP) HUC represents the 4 LHC experiments at CERN that fully rely on the use of **grid** computing for their offline data distribution, processing and analysis. Increasing focus is placed on common tools and solutions across these four large communities together with their re-use by other HEP experiments as well as numerous different disciplines and projects. Areas supported include: Data Management, Data Analysis, Persistency Framework and Monitoring
Life Sciences	The Life Science (LS) HUC originates from the use of **grid** technology in the medical, biomedical and bioinformatics sectors in order to connect worldwide laboratories, share resources and ease the access to data in a secure and confidential way through the health-grids.
Astronomy and Astrophysics	The A&A HUC is devoted to the evaluation of different solutions for the gridification of a rich variety of applications, as well as the accomplishment of a good level of interactivity among different technologies related to supercomputing, i.e. High Performance and High Throughput Computing, **grid** and cloud.
Earth Sciences	Earth Science (ES) applications cover various disciplines like seismology, atmospheric modelling, meteorological forecasting, flood forecasting and many others. Their presence in SA3 is currently centred in the implementation, deployment and maintenance of the EGDR service to provide access from the **grid** to resources within the Ground European Network for Earth Science Interoperations - Digital Repositories (GENESI-DR). The ES HUC includes also researchers and scientists working in the climate change domain. In particular most of them actively participate in the Climate-G use case. This use case exploits the GRelC service for distributed metadata management and the Climate-G portal as scientific gateway for this collaboration.

EGI Collaboration



European Grid Infrastructure

(April 2011 and yearly increase)

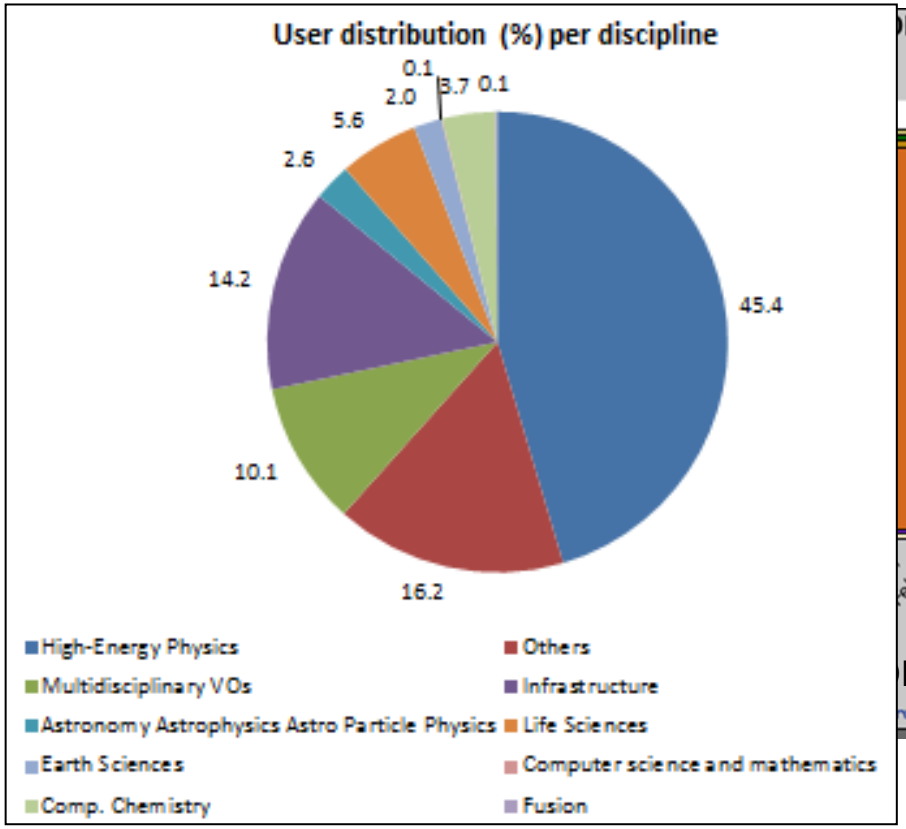


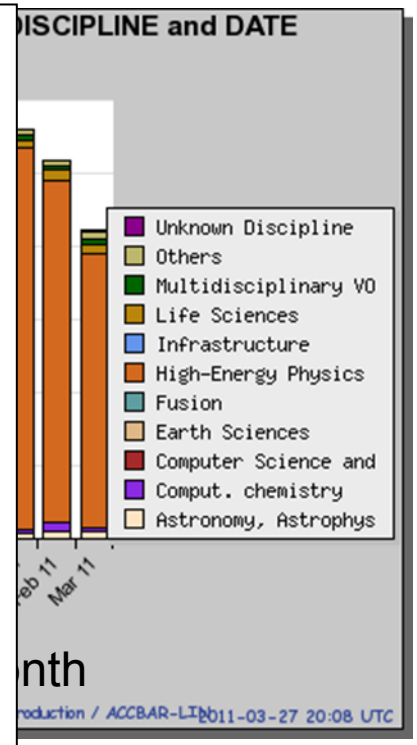
EGI Usage (April 2011)

11226 End-users (-7.7%)
219 VOs (+17.7%)
~30 active VOs: constant

User Communities

- Archeology
- Astronomy
- Astrophysics
- Civil Protection
- Comp. Chemistry
- Earth Sciences
- Finance
- Fusion
- Geophysics
- High Energy Physics
- Life Sciences
- Multimedia
- Material Sciences
- …





Average usage 2010-2011 vs 2009-2010

- 26.6M jobs/month, 873,400 jobs/day (+420%)
- 74.6M CPU wall clock hours/month (+86.5%)
- 551M HEP-SPEC06 CPU wall clock hours/month (+99.4%)

Specific Achievements

- In its first year, EGI-InSPIRE has shown that it is possible to identify and provide common solutions across multiple communities – even during the production phase (of the LHC)
- Includes work on:
  - Data Management
  - Catalogue / Storage Element Consistency
  - Data Placement / Dynamic Caching
- At conceptual, architectural, implementation and even deployment level

Common Solutions - Examples

Tool / Service	Description	Communities
Ganga	Extensively used as a "gridification tool" by many projects / disciplines. This includes not only communities within EGI-InSPIRE but also others in many fields	Numerous non-HEP and non-HUC projects
Mini-Dashboard	To be used with Ganga to monitor Ganga-based activity on the grid. Used by EnviroGRIDS and offered to NA3	EnviroGRIDS, NA3
Experiment Dashboards	Common schema and code base for all job monitoring applications: job summary, historical view & task monitoring – implemented from July 2010	ATLAS + CMS
GRelC	Uniform access to relational DBs and flat files via portal	LS, A&A
MPI	High impact on multiple user communities	CCMST, A&A, F
Frameworks	Use of LHCb's DIRAC framework by LCD/ILC and Belle collaborations. Investigation of DIRAC by Earth Science and other communities	HEP (beyond LHC), ES
Data Management	Data popularity (dynamic data placement / caching), consistency of catalogs / storage	LHC VOs
Site Stress Testing	HammerCloud service used at ATLAS, CMS and LHCb. Fully applicable to other VOs / communities	HEP

Data Management

- Crucial area for LHC and other HEP experiments
  - Data volumes: tens of PB/year, rates: up to 200TB/day between sites, several hundred active analysis users / experiment, 1M analysis jobs / day
- Experience from first data taking has shown that some assumptions are no longer optimal
  - Based on decade old model “MONARC” which assumed network was scarce resource – more in talk from Ian Fisk
- Role of tape: increasingly used only as archive rather than active data store as in past HEP experiments
  - Move to decoupled but more predictable storage for Disk / Archive areas
  - Eases operation and development

Data Placement & Dynamic Caching

- Initial Phase of HEP Data Distribution was based on static pre-placement
  - Significant fraction of such data never read!
- Computing Models now driving towards dynamic data placement
  - Replication is based on usage (“popularity”) – this results in better network and storage utilization
- Implemented first for ATLAS, now for CMS and LHCb

Catalogue Consistency

- With 50PB of data storage across a large number of sites worldwide, inconsistencies can easily arise!
  - Data that resides on Storage Elements but not in various catalogs (grid, experiment) referred to as “Dark Data”
    - One site recently reported 70TB dark data!
- Using a messaging-based system, various catalogs and SEs can talk to each other and implement lazy synchronization

The Next Challenge

- EGI-InSPIRE has demonstrated that multiple disciplines can work together, which leads naturally into the next challenge
  - It is possible to identify **common requirements** in the domain of Massive Scale Data Management that cross a wide variety of domains / disciplines and address “the Digital Agenda”
  - A key area to be supported in the next round of EU projects – from 2014 but with a long-term vision: 2020+

Digital Agenda – Overview

- The EU intends to invest significantly in solutions that enable all data-driven applications – the entire data life-cycle
- It wishes to support solutions that address as wide a range of disciplines as possible – from science to humanities and beyond
  - eHealth, climate change, science and education explicitly mentioned amongst many others
- As a multi-disciplinary project that is largely data-driven, EGI-InSPIRE offers an attractive foundation on which to build such an effort

Requirements Summary

- Data preservation: up to millennia
- Data volumes: 100PB and growing
- Data access: need both coarse and fine grain access with flexible and scalable authentication and authorization

- Solutions that meet the combined needs of these communities are likely to be able to handle those of many others
- Natural sequitur: prove it → service

Summary

- EGI-InSPIRE is a multi-disciplinary project that requires us to look for common, sustainable solutions
- Building a matrix of common requirements for data management is an important step in preparing for the next generation of funding
- Solutions which can address requirements spanning not only multiple scientific disciplines but also applicable to a wider range of communities will be extremely valuable in the future
- A key area of the next round of EU funding – FP8