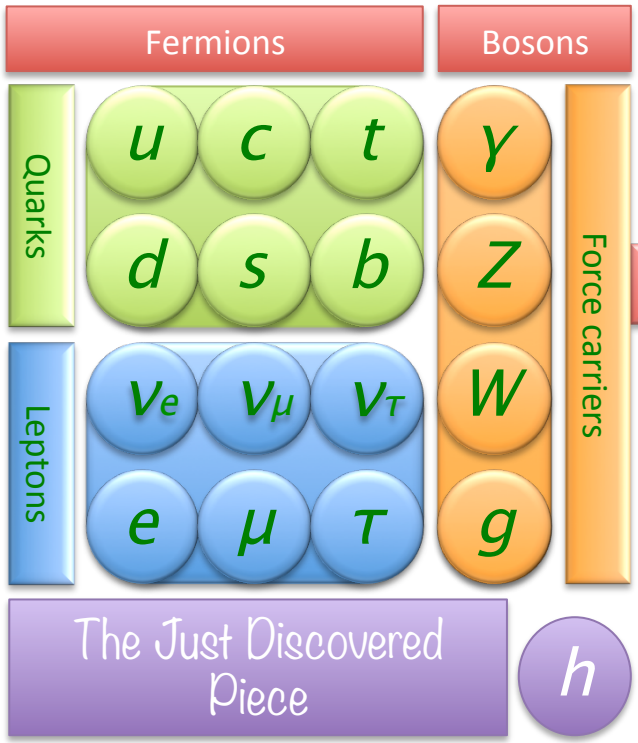
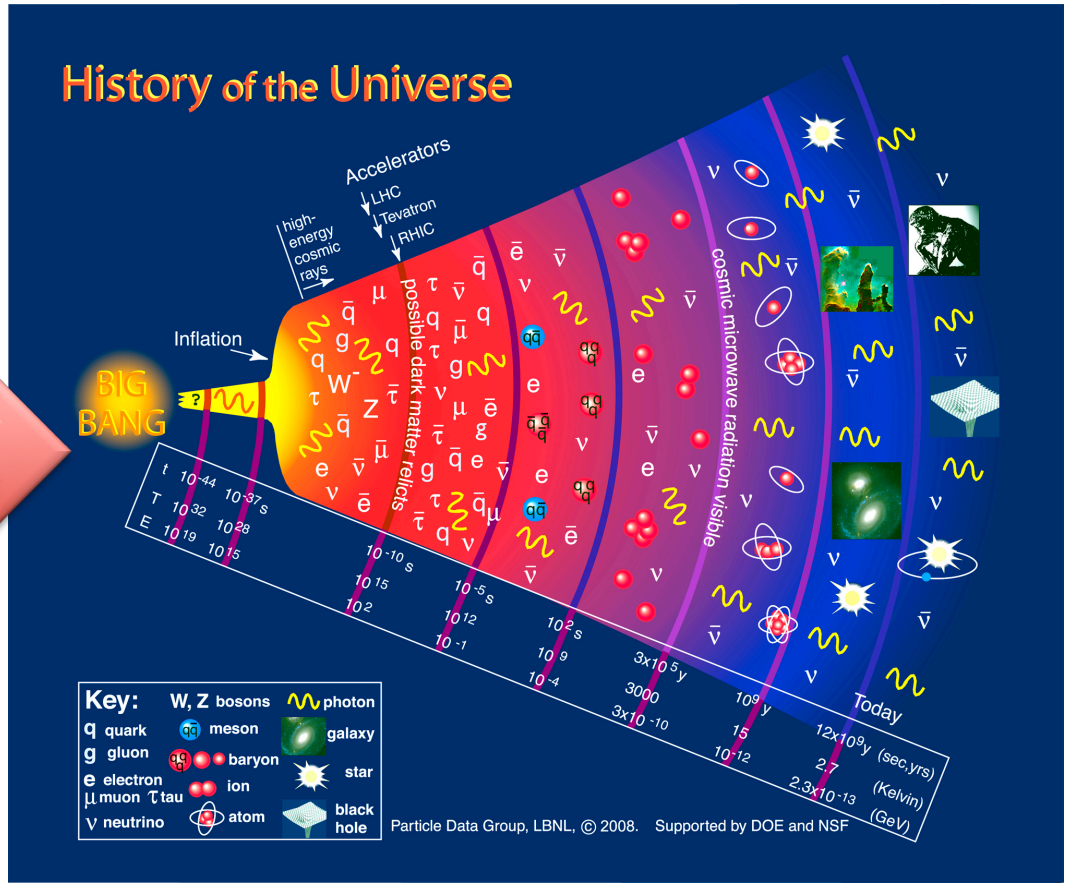
Open Data & Data Management: the INFN experience

Marcello Maggi
INFN Senior Researcher
Istituto Nazionale Fisica Nucleare
Bari-Italy

A Chaotic view
from a scientist point of view

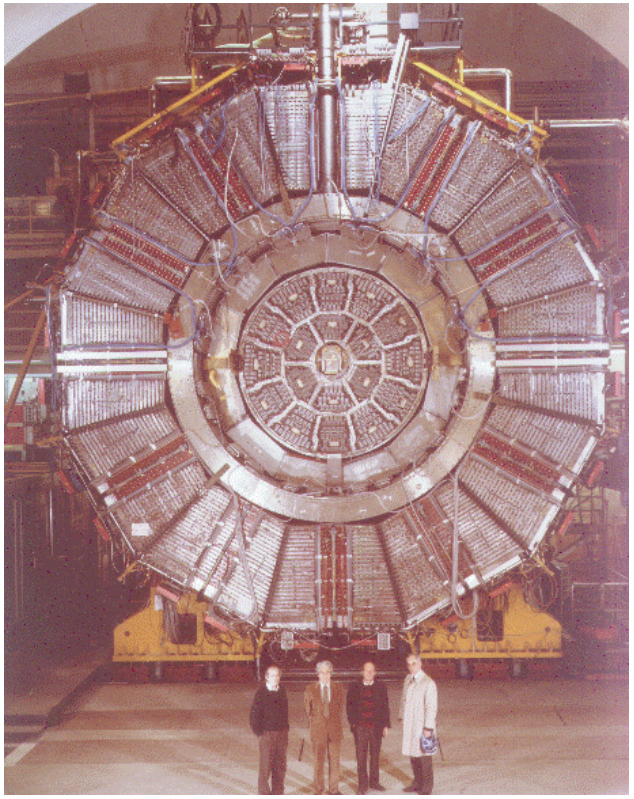
The HEP Scientist



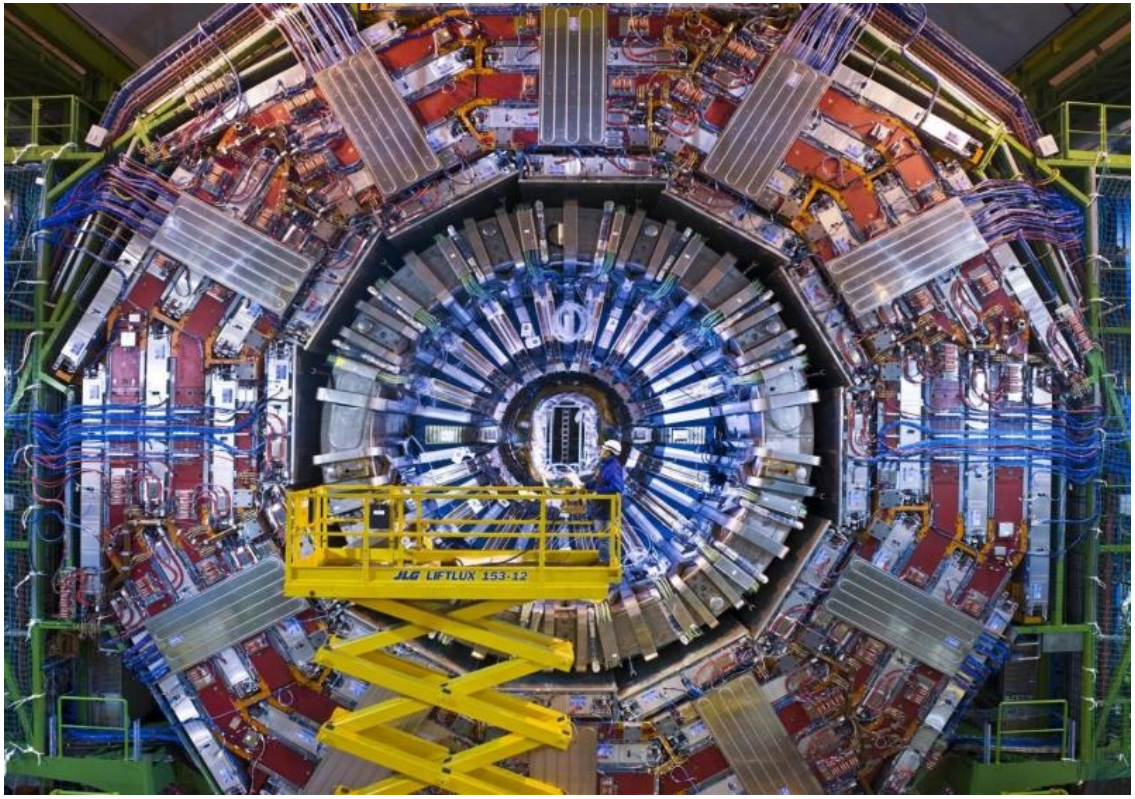


FROM MICROCOSM → Dark Matter
Matter/Anti-matter Asymmetry
Super Symmetric Particles → TO MACROCOSM

In Big Communities
In International Labs (CERN)



Past Century collaboration
~500 Scientists

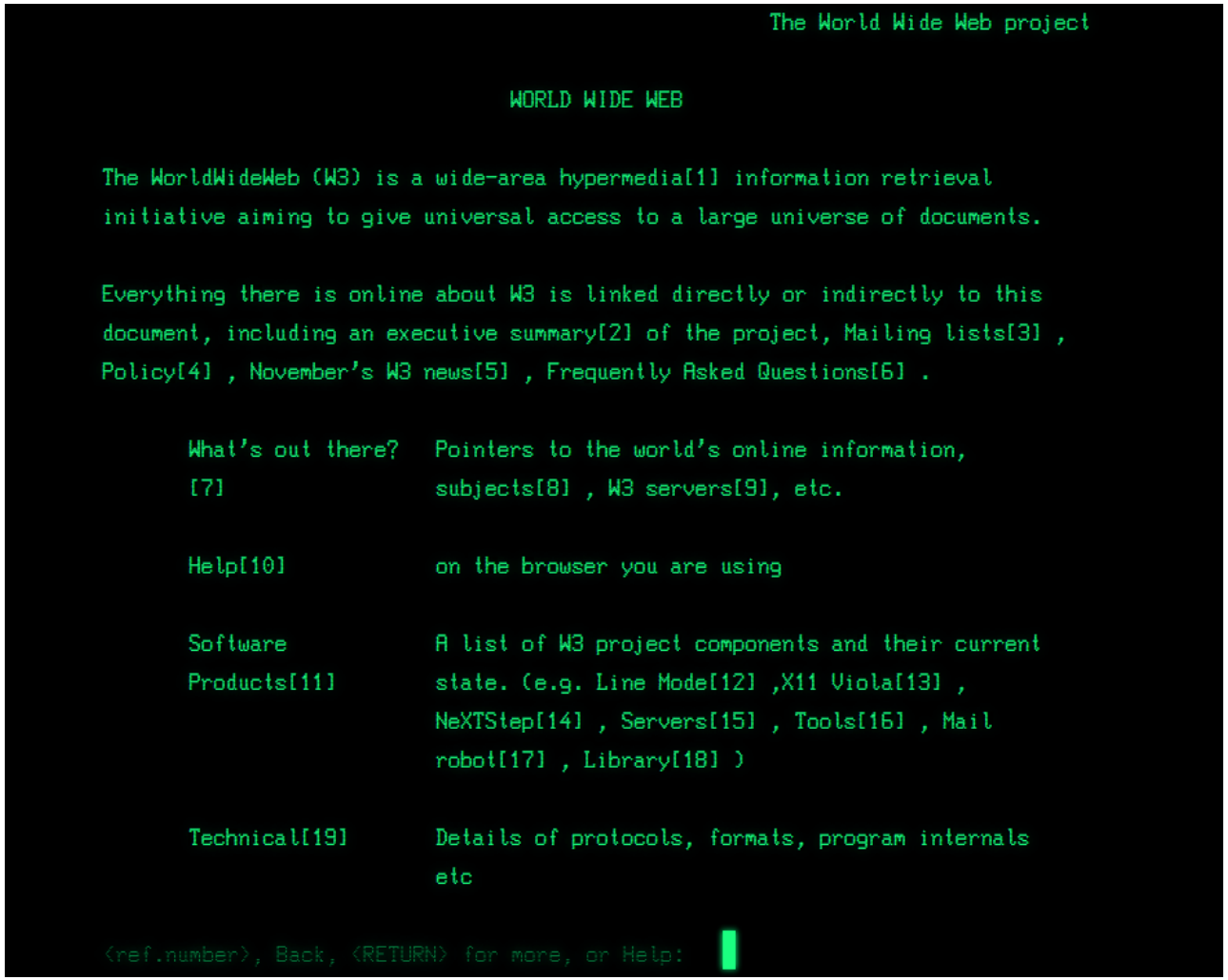


Today collaboration
~4000 Scientists

From all around the word

Data Sharing &
Data Management
Fundamental Issue

Birth of Web @ CERN



INFN

- Community of researcher in physics and applied physics
- Based on 4 national laboratories and 20 divisions spread across Italy

Big impact on Italian Society

The Italian e-Infrastructure

Today Picture:

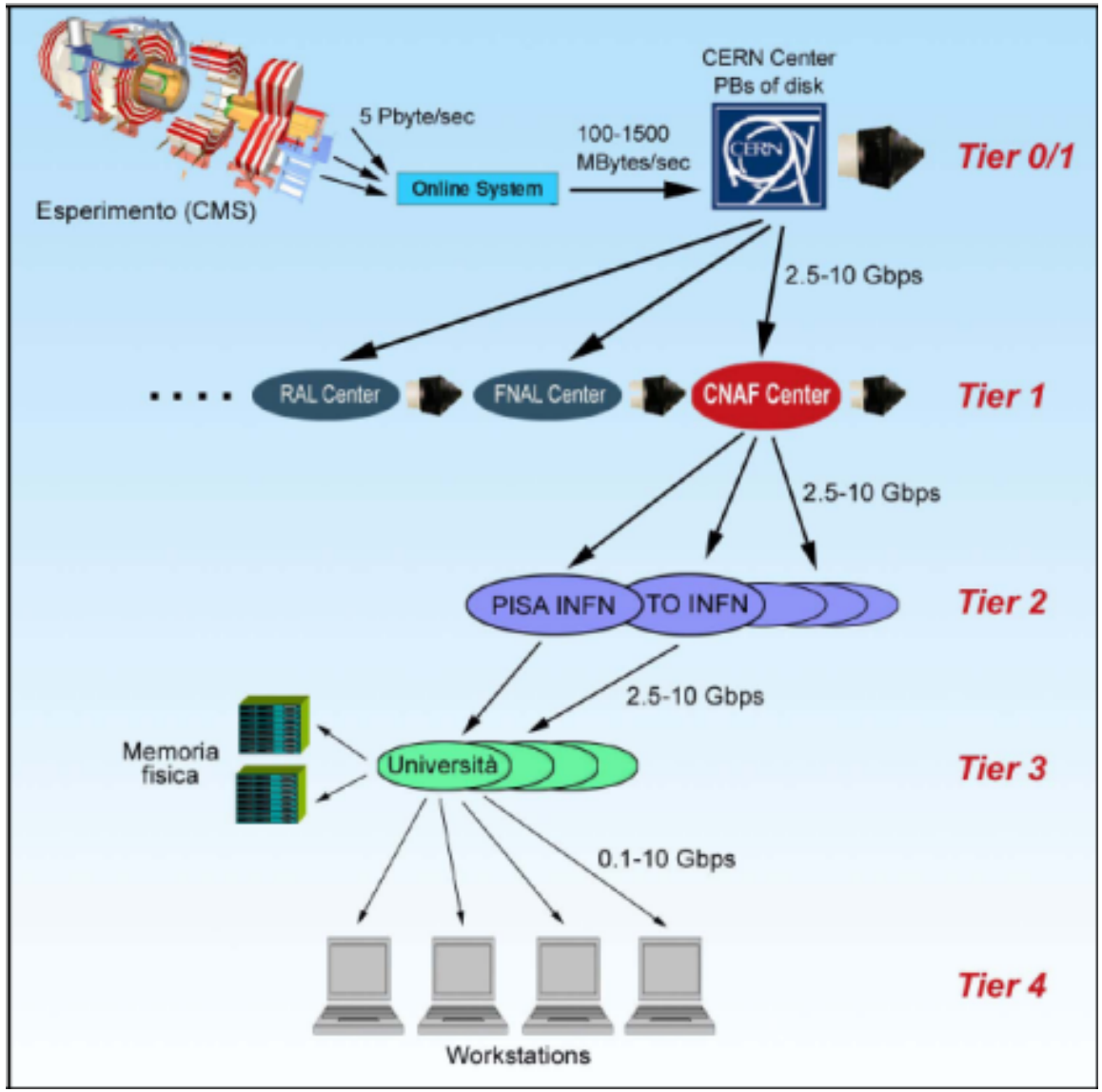
50 data centers

40,000 cores

~60 PB

Growing through approved projects

- CPU: +25%/year
- Disk: +20%/year

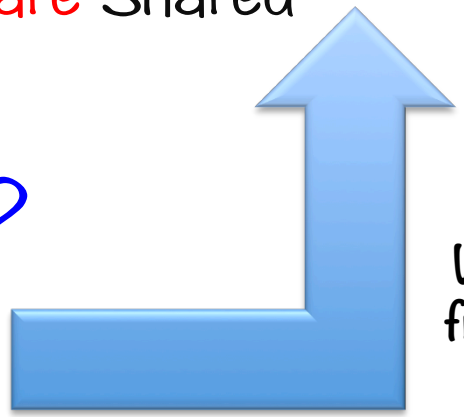


The (Big) DATA

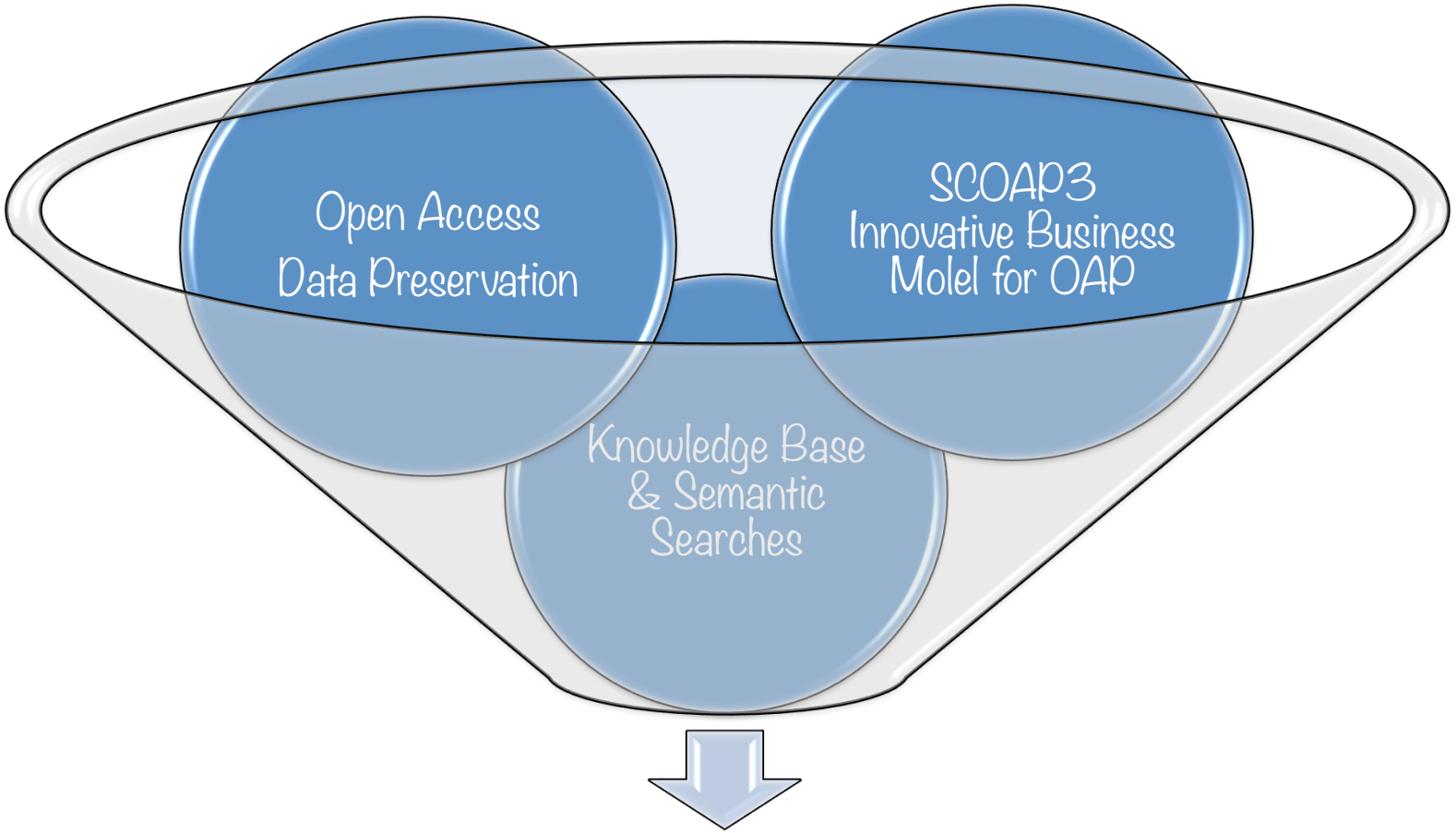
10^7 "sensors" produce 5 PByte/sec
Complexity reduced by a Data Model
Analytics in real time filters to 0.1–1 Gbyte/sec (Trigger)

Data + Replica move with a Data Management Policy
Analytics produce "Publication Data" that are Shared
Finally the Publications

Is all that Open?



Open Science



Common Practices

INFN & Open Access

Budapest Open Access Initiative 2001

Berlin Declaration on **Open Access** to Knowledge in the Sciences and Humanities 2003

INFN in SCOAP3 2007

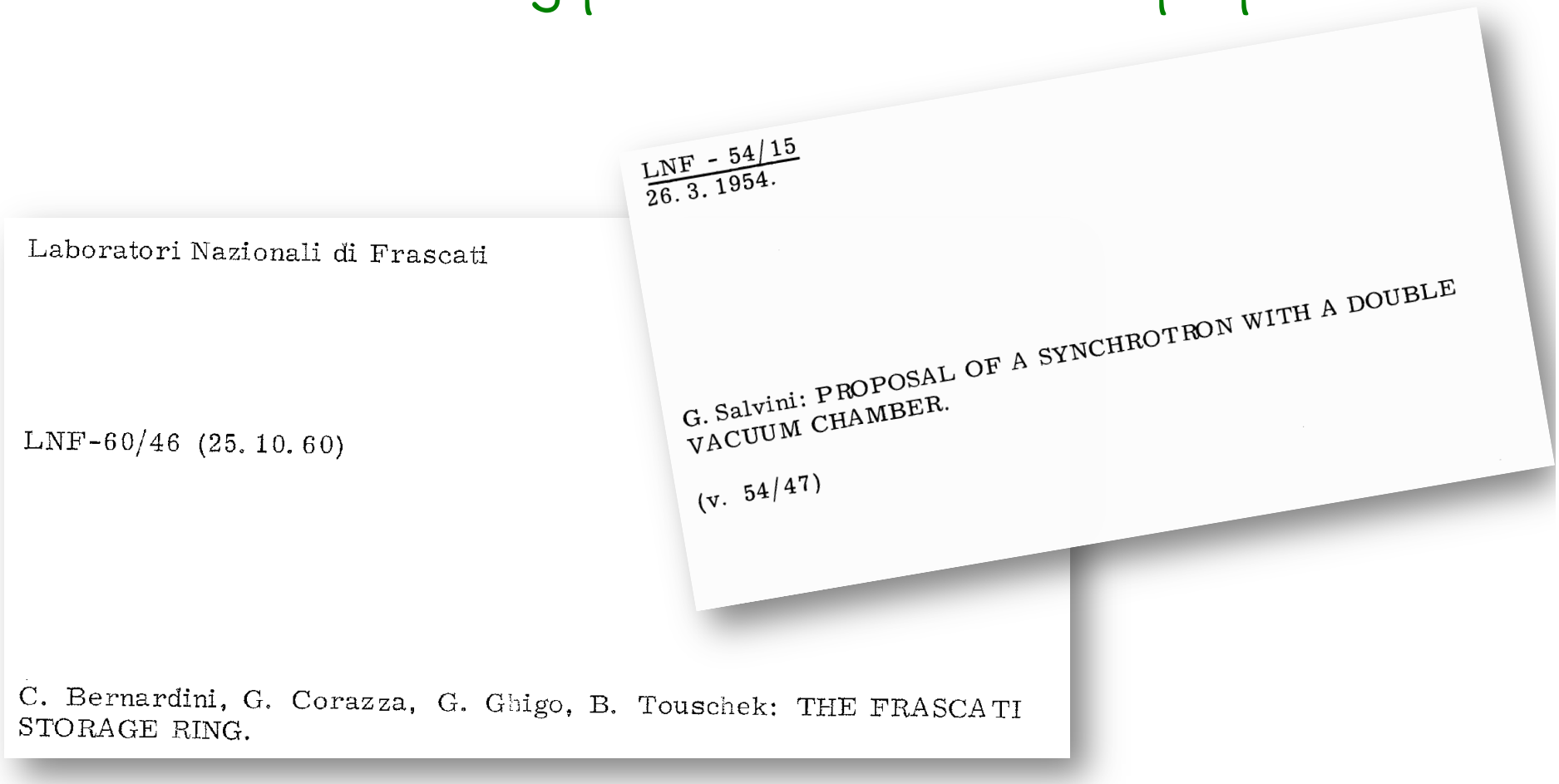
INFN signs Berlin Declaration 2008

INFN signs Granada Declaration 2010

INFN signs the MedOAnet *position paper* 2013

Since Ever

HEP community publish and distribute preprints



LNF - 54/15
26. 3. 1954.

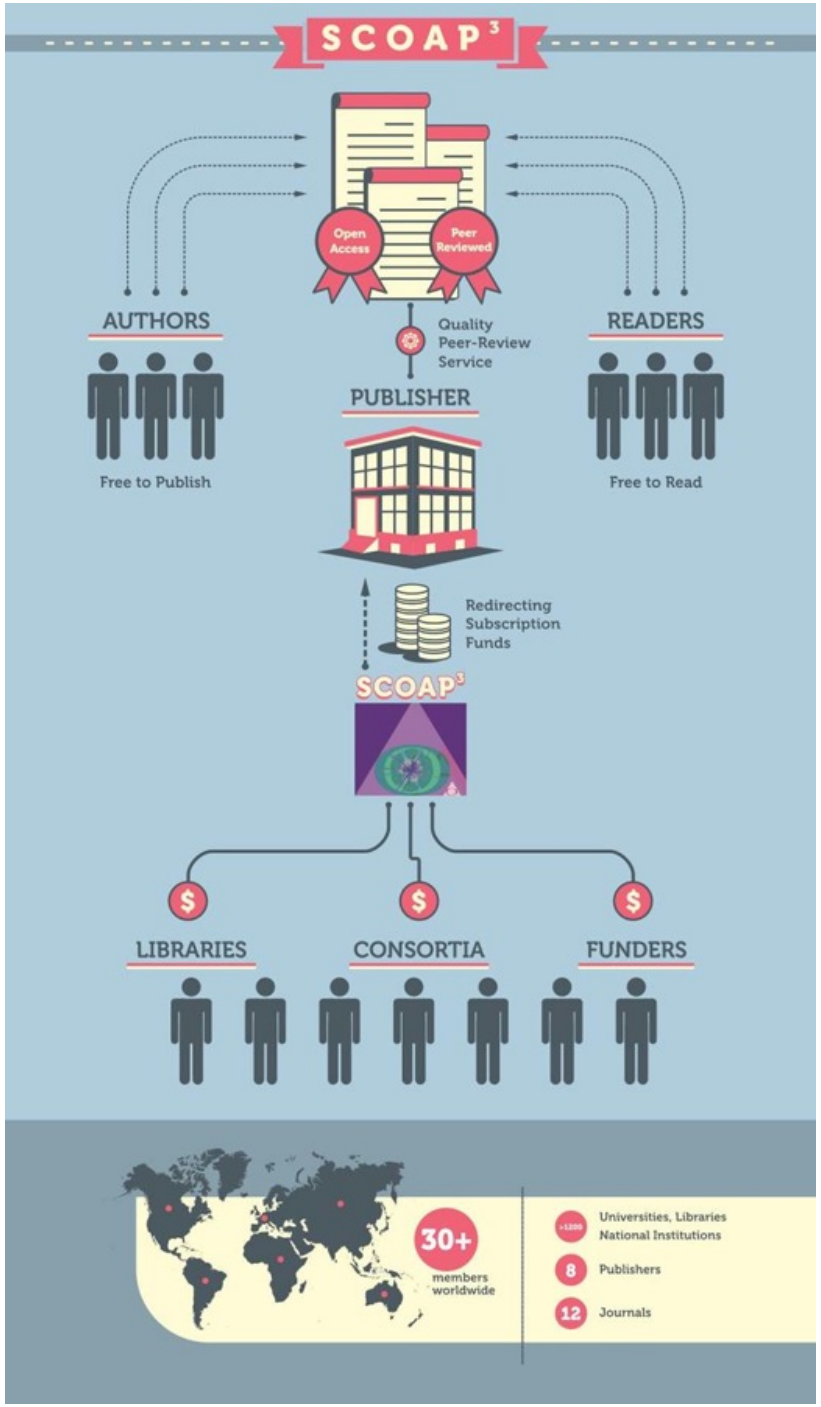
G. Salvini: PROPOSAL OF A SYNCHROTRON WITH A DOUBLE VACUUM CHAMBER.

(v. 54/47)

Laboratori Nazionali di Frascati

LNF-60/46 (25. 10. 60)

C. Bernardini, G. Corazza, G. Ghigo, B. Touschek: THE FRASCATI STORAGE RING.

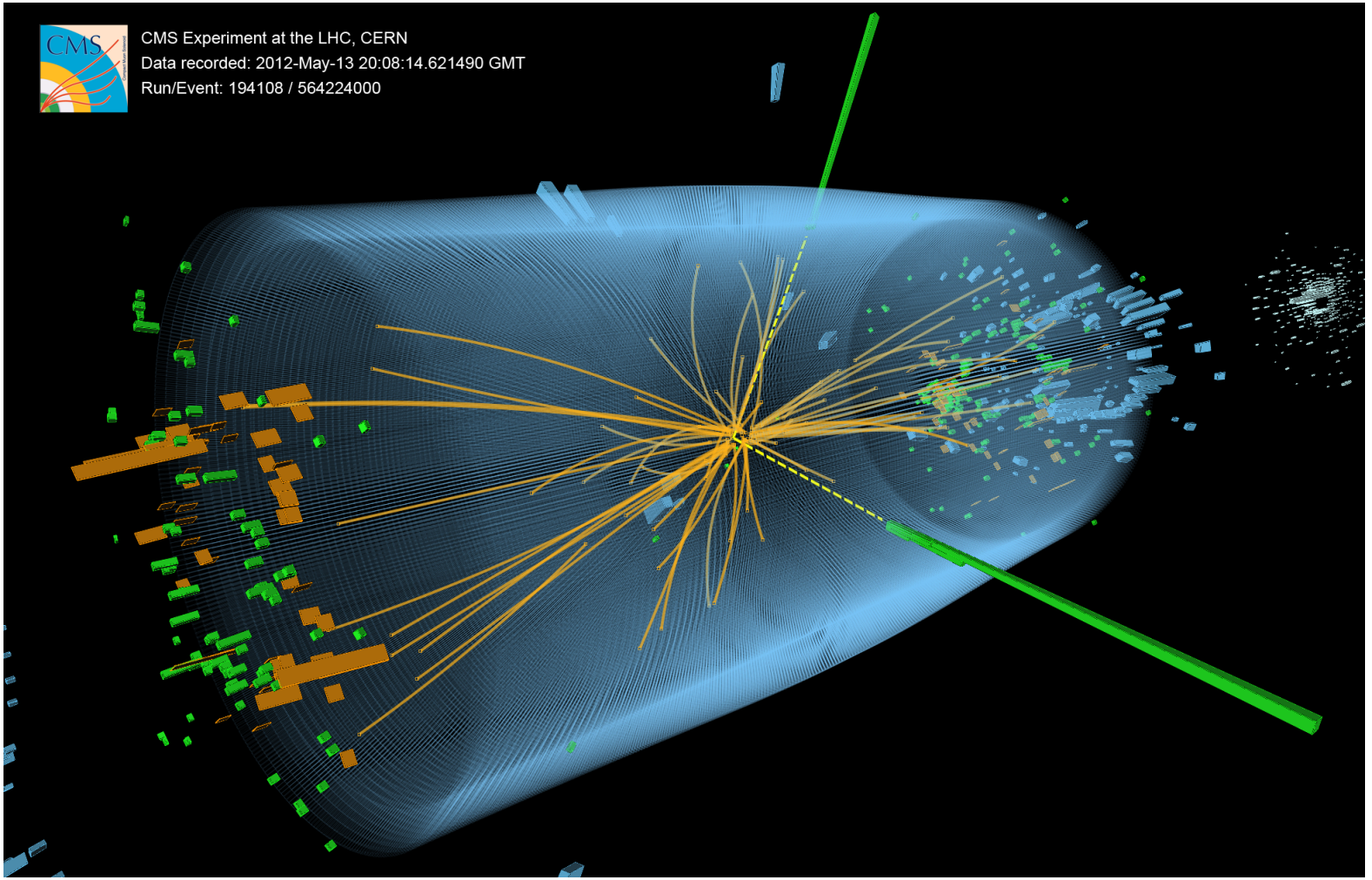


Worldwide consortium funding HEP publications and enforce OA, through the re-routing of subscription funds, and the transition to a system of commercial competition

Past: The HEP agencies subscribing through libraries funded peer-reviews, allowing its users to read the articles. There was no form of commercial competition between journals.

Present: The HEP agencies and libraries, together, contribute to the consortium SCOAP3 that, after selecting the journals, pay centrally the peer review for each published article. The articles are OA.

Open Data



Event Display of Higgs boson decay

Publication Data

Analytic step 1

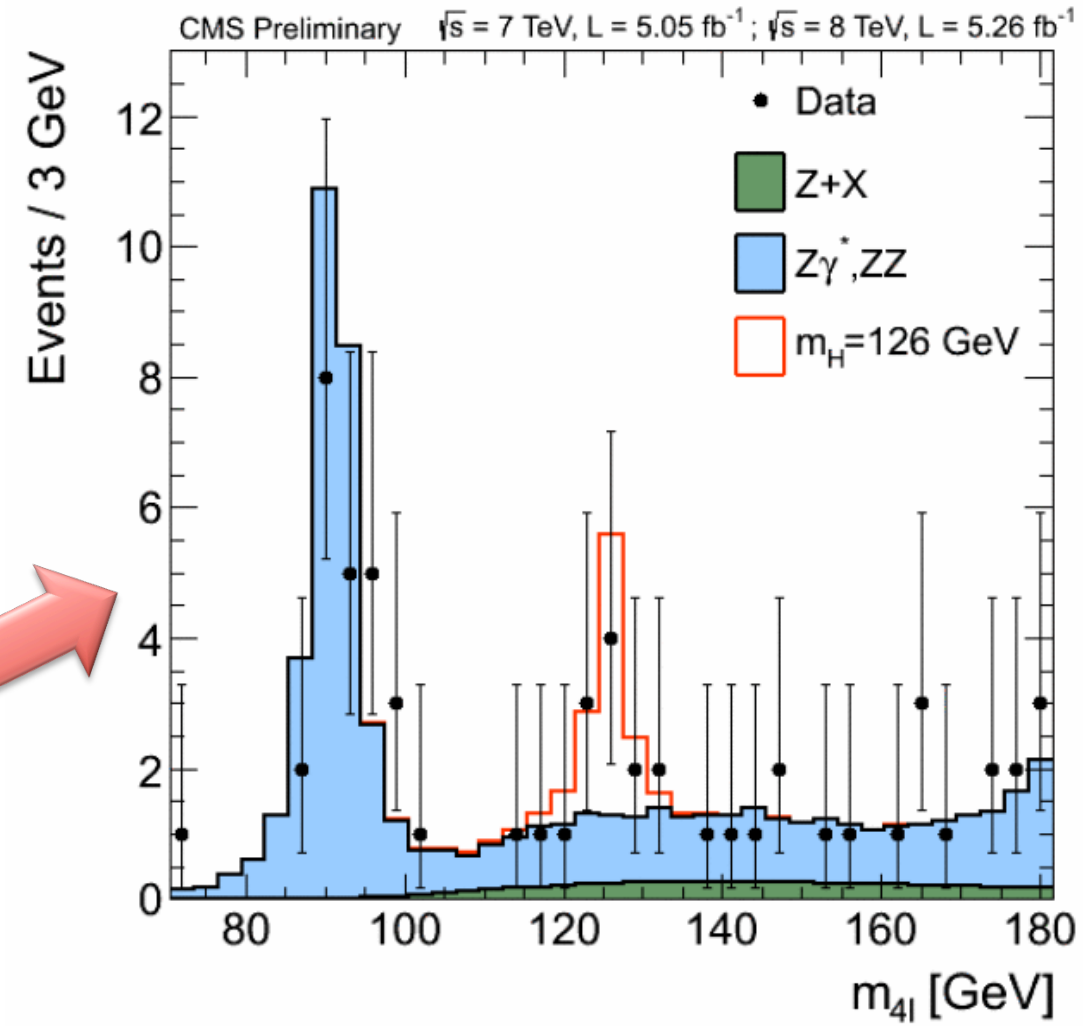
Pre Selected Data

Analytic step 2

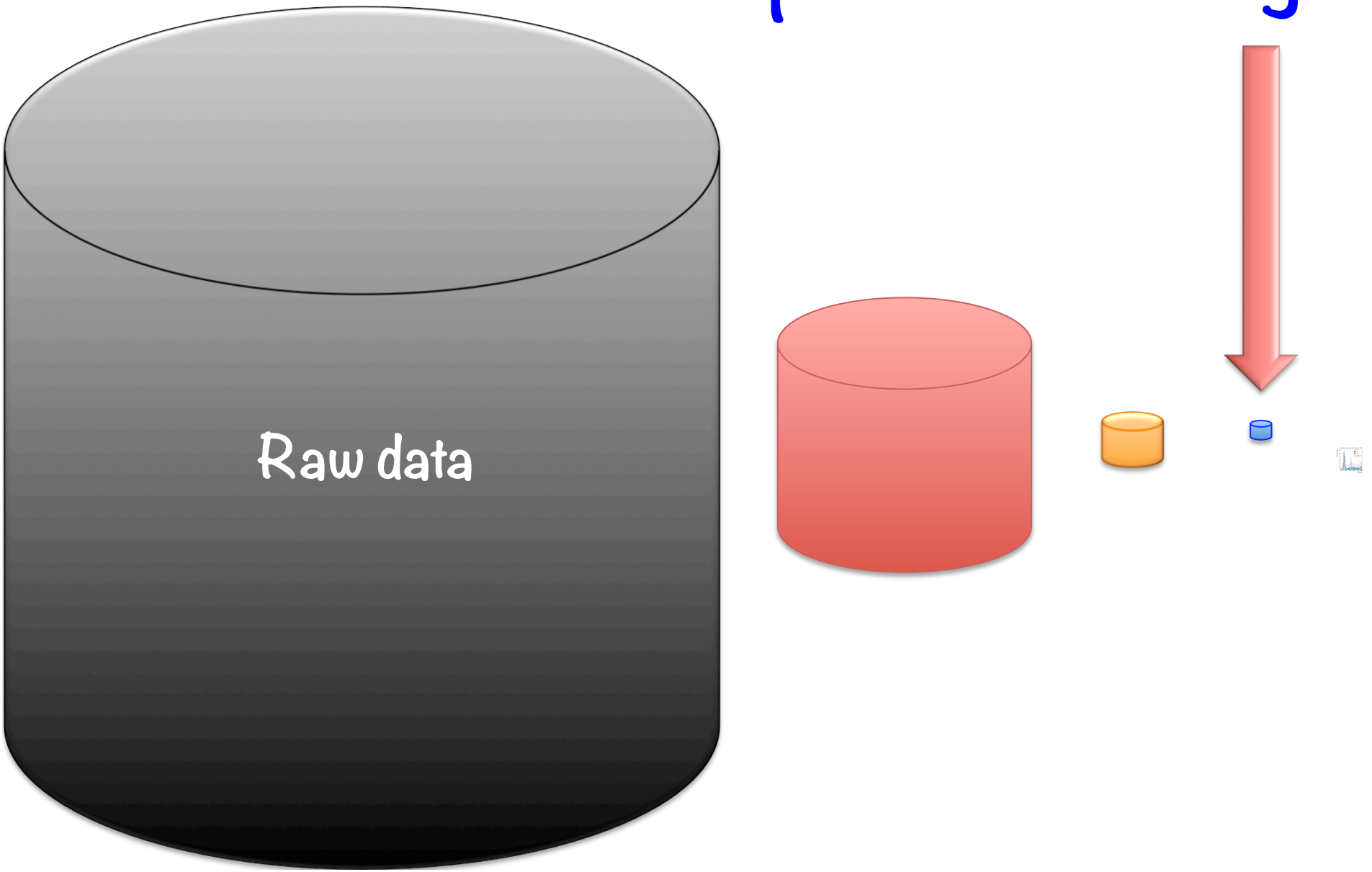
Final Data Samples

Analytic step 3
Analytic step 4

...



The tip of the iceberg



Levels of Open Data? Discussion on going

Data Harmonization Guidelines

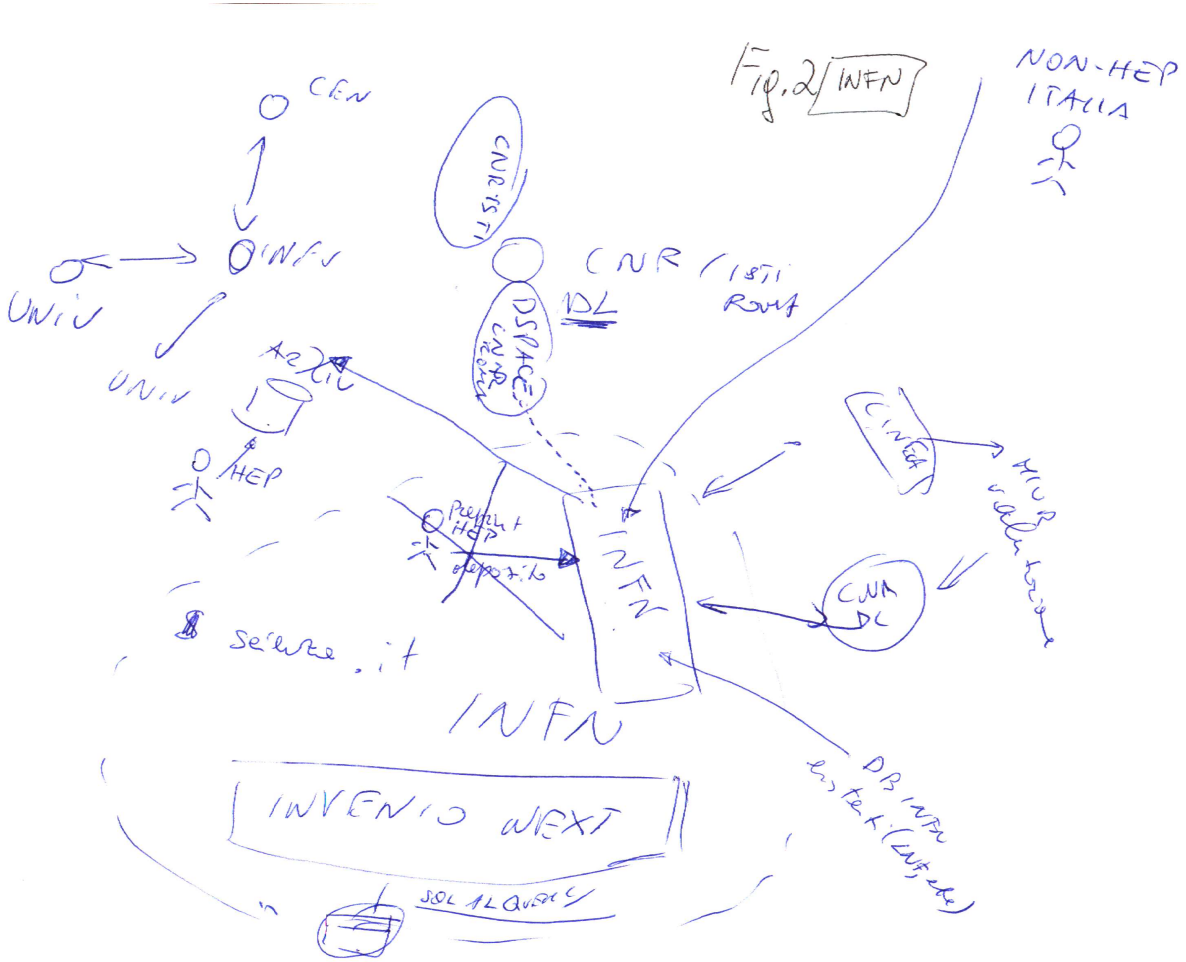
- Common tacit points of agreement between LHC experiments:
  - ✔ **level 1 data:** All experiments already make data from papers and supporting information available through HEPDATA/Inspire, support open access journals etc..
  - ✔ **level 2 data:** All experiments already support limited access of samples in simple formats for outreach and teaching.
  - ! **level 3 data:** Full reconstruction outputs for analysis (AOD, DPD/ntuples) might be made available after an embargo period – but suggested durations range from 3 to 10 years, and there is a question of usefullness. The resource implications to make this useful are high.
  - ✔ **level 4 data:** General agreement RAW data is preserved for the experiment and future – open data access is not usually possible even to the collaboration members. (In ATLAS access to RAW data on tape is restricted).

- **Tools like Rivet, HEPDATA & Recast may make data (information) usefully available, bridging level 3 and level 1.**

INFN Open Data

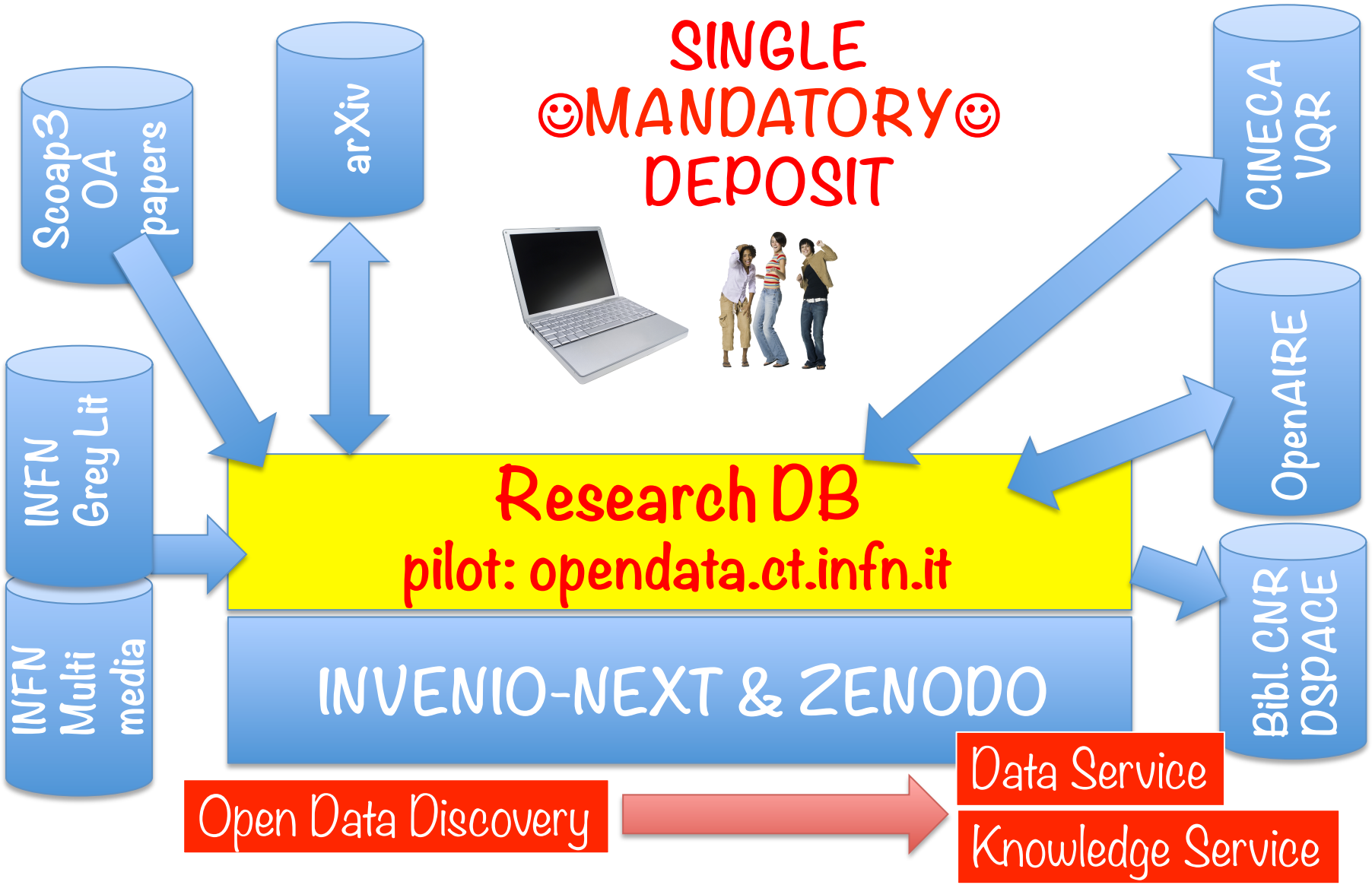
PILOT SCREENSHOT
opendata.ct.infn.it

Italian Research DB

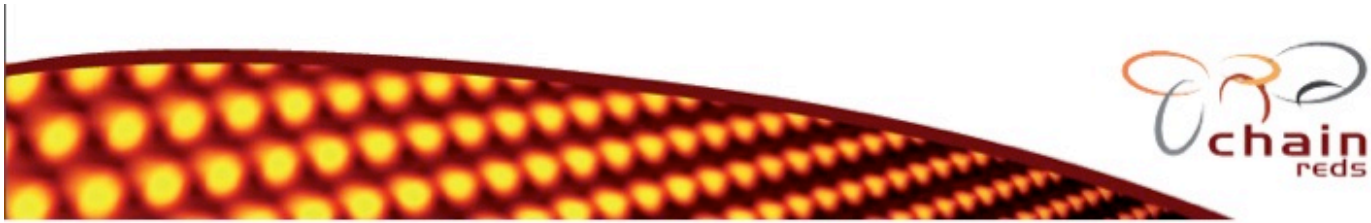


Resulting from a discussion between the CERN and INFN responsible persons for Open Access

Happy INFN scientists



INFN is Active in Knowledge Base & Semantic Search



Co-ordination & Harmonisation of Advanced e-Infrastructures for Research and Education Data Sharing

The CHAIN-REDS Knowledge Base

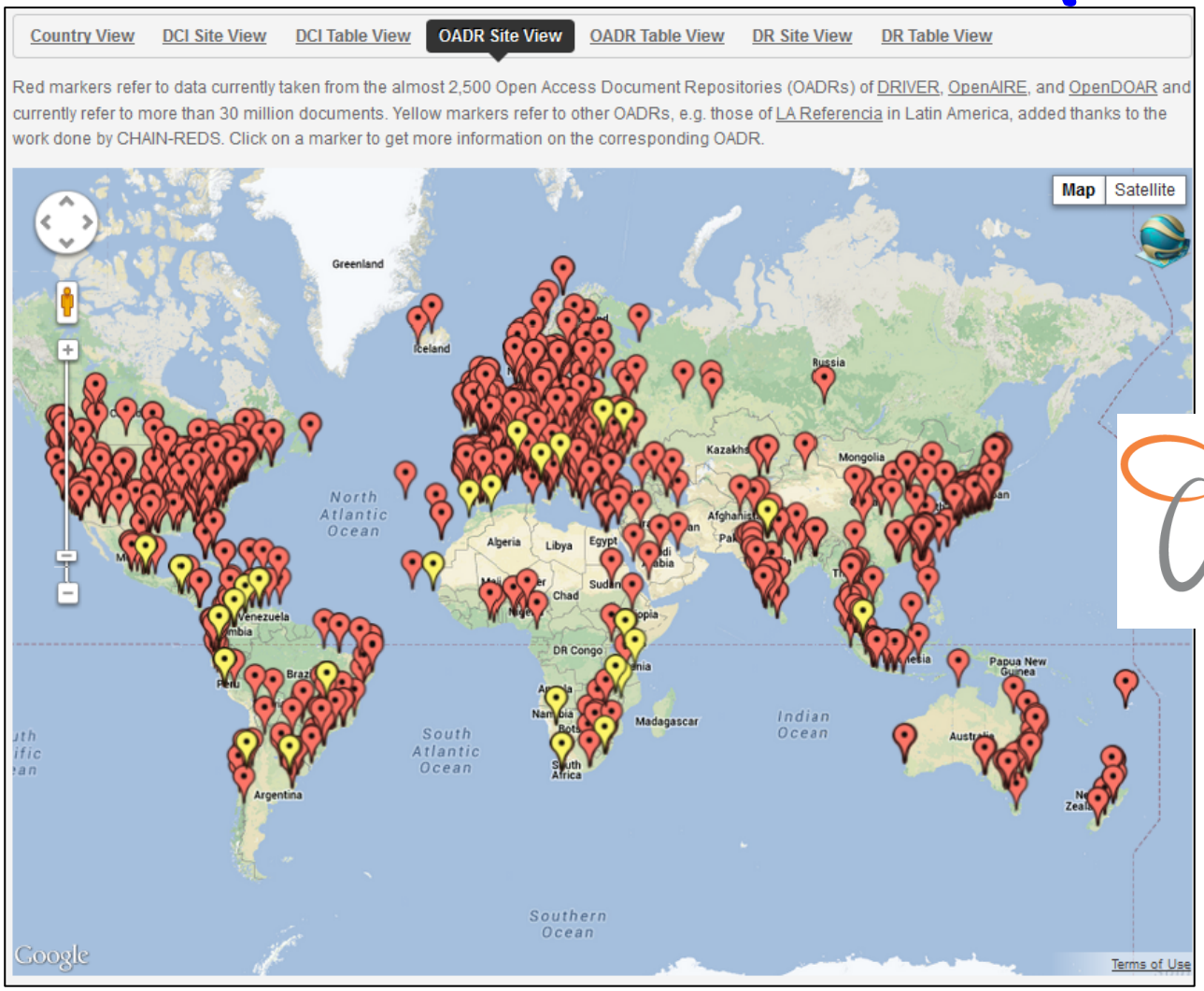
Roberto Barbera – University of Catania and INFN - Italy
(roberto.barbera@ct.infn.it)

Research Infrastructures – Proposal n. 306819

www.chain-project.eu
proj-office@chain-project.eu

e-infrastructure

A Global OA Repository



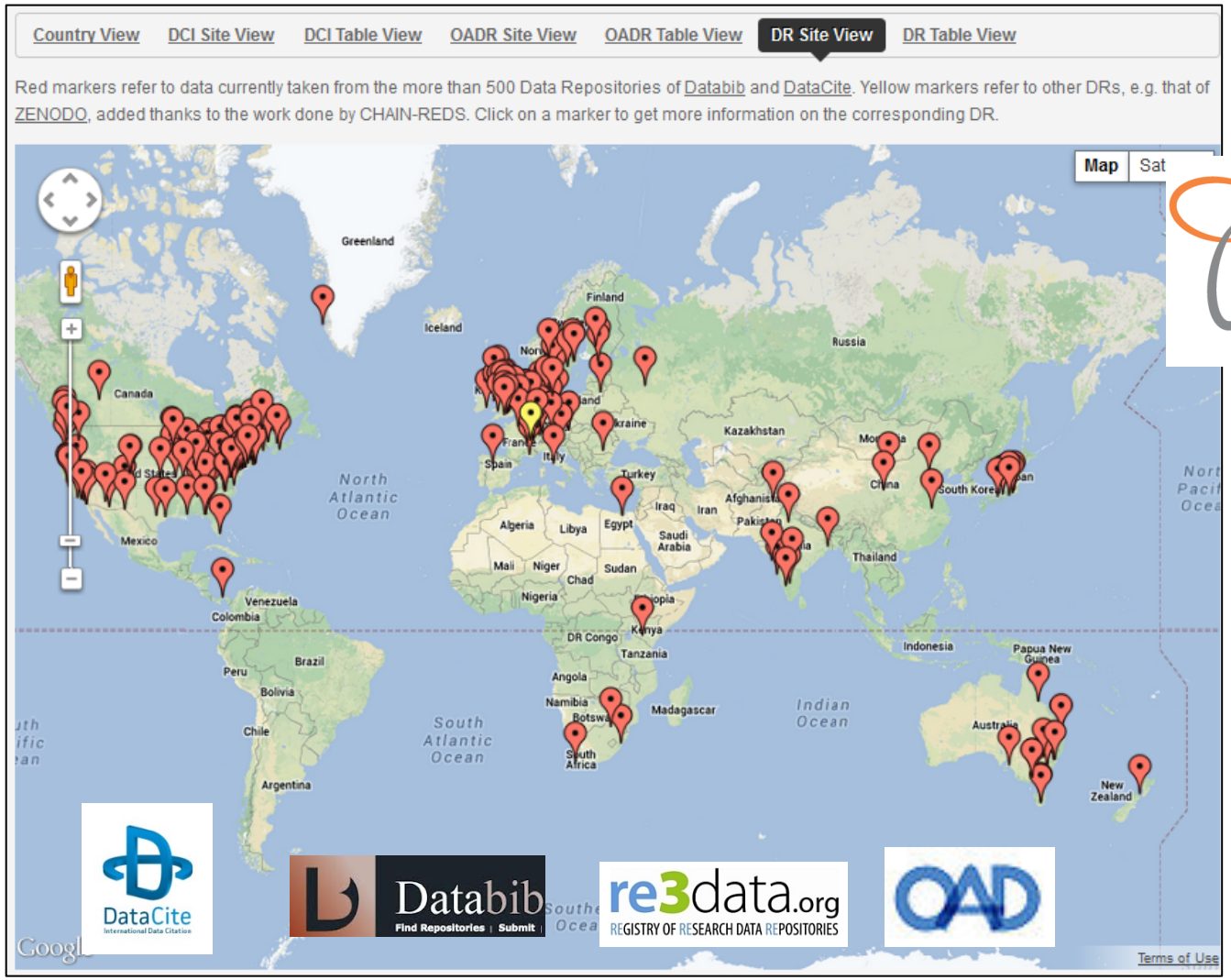
~2,500 repos
>33 M docs

chain reds

Registry of Open Access Repositories (ROAR)

driver OpenAIRE OpenDOAR

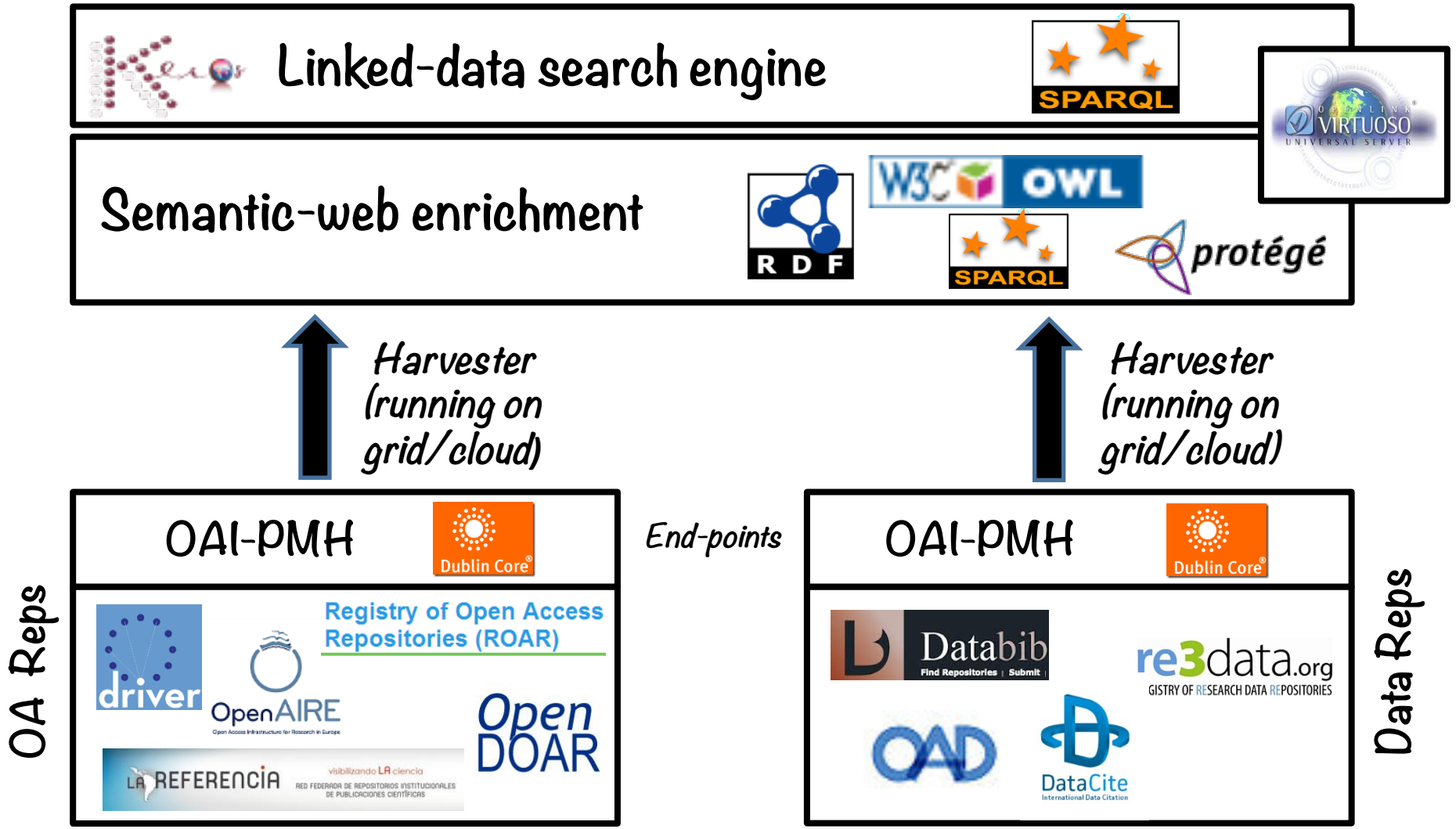
Global Data Repository



chain reds

~600 repos
Lots of data !

Data & Knowledge Infrastructure



European Research e-Infrastructures

New Trend in Europe:

Secure computing resources funding from FA:

- ELIXIR (Life science) identified nodes in the consortium
- LifeWatch (Earth science) has IT research center
- CLARIN (Arts, humanities and social science) has certified centers

Virtual hubs federating major computing centers to offer resources and services

Eu-TO

Federate major computing and data process centers of Particle, Nuclear, Astro-Particle Physics, Cosmology and Astrophysics into a integrated distributed infrastructure:

a virtual European TierO data and computing center around which all other national centers revolve and from which all concerned national e-infrastructures radiate

IN2P3-Fr INFN-It STFC-UK DESY-DE

KIT-DE IFAE-ES CIEMAT-ES CERN

signed the position paper

NeIC (Nordic e-Infrastructure Collaboration) asked to join

INFN is exporting/importing experience

Multidisciplinary and/or extra Europe

- Chain-Reds (Coordination and harmonisation of e-infrastructure for research and data sharing
- agINFRA (data infrastructure for agriculture)
- DCH-RP (Digital Cultural Heritage Roadmap for Preservation)
- BioVel (Biodiversity Virtual E-Laboratory)

National collaborations on

- Computational Chemistry: Uni. Pg, Uni. To, CNR-ISOF
- Environmental Science: EMSO (European Multidisciplinary Seafloor Observatory) (ESFRI), DRHIM (Distributed Research Infrastructure for Hydro-Meteorology) (FP7 proj.) CIMA (Centro Monitoraggio Ambiente)
- Bioinformatics: CNR-ITB, Uni. Bo

Partecipazione JRU

- Elixir (European life science infrastructure for Biological Information)
- Life Watch : earth science in progress

From Global to Local Projects

- "Core Business" Projects
  - DHTCS-it
  - ReCaS
  - Prin-Stoa
- Multidisciplinary Projects (smart cities)
  - Prisma (PiattafoRme cloud Interoperabili per SMArt-government)
  - OCP (Open City Platform)
  - Cagliari 2020

Open Cloud Platform -1

Partners

1. Almaviva the Italian Innovation Company S.P.A
2. Maggioli SpA
3. Santer Reply S.P.A.
4. Pluservice s.r.l capofila della ATI Marche (E-LINKING ONLINE SYSTEMS S.R.L., ETT S.p.A., FILIPPETTI S.P.A., APRA PROGETTI S.R.L., HALLEY INFORMATICA S.R.L., ESALAB S.R.L., SEDA S.p.A. - Gruppo KGS, ITALSOFT S.R.L., JEF S.R.L.)
5. LASCAUX s.r.l. capofila della ATI Toscana-ER (SISTEMI TERRITORIALI S.R.L., SINED S.R.L., PHOOPS S.R.L., AGENZIA ESPRESSI S.A.S., 3D INFORMATICA S.R.L.)
7. INFN - Istituto Nazionale di Fisica Nucleare
8. UniCam - Università degli Studi di Camerino

Open Cloud Platform -2

PA involved

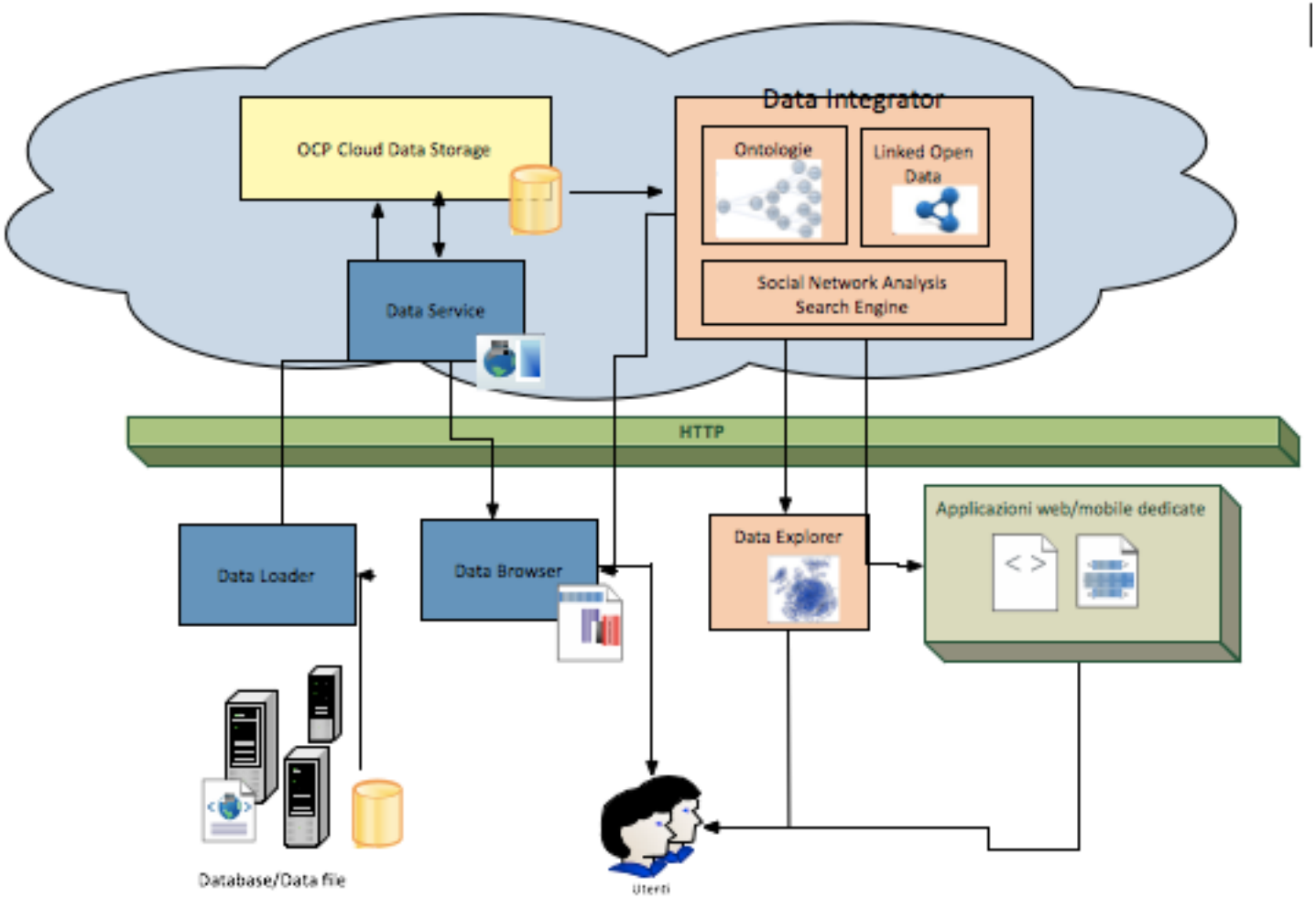
ITALIAN REGIONS

1. REGIONE MARCHE
2. REGIONE TOSCANA
3. REGIONE EMILIA ROMAGNA

COMUNI/UNIONI

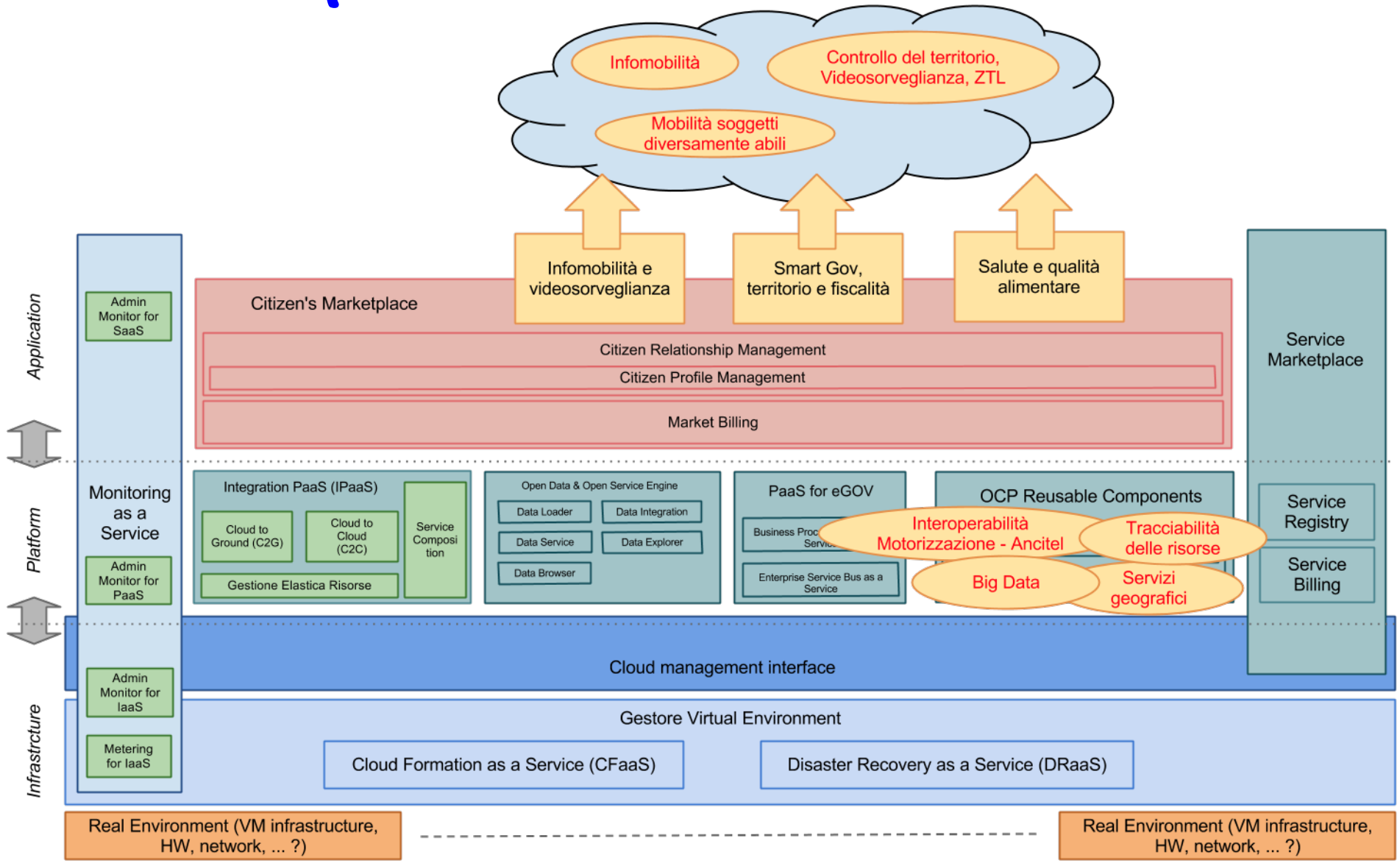
1. Comune di Macerata
2. Comune di San Severino
3. Comune di Camerino
4. Comune di Matelica
5. Comune di Castelraimondo
6. Comune di Tolentino
7. Comune di San Benedetto
8. Comune di Ancona
9. Comune di Pesaro
10. Comune di Senigallia
11. Comune di Civitanova
12. Comune di Fabriano
13. Comunità Montana Alto e Medio Metauro
14. Comune di Ascoli
15. Comune di Rosignano Marittimo
16. Comune di Livorno
17. Comune di Lucca
18. Comune di Massa
19. Unione dei Comuni dell'Amiata Grossetana
20. Comune di Cesena
21. Unione dei Comuni della Bassa Romagna

Open Cloud Platform -3



Open Data & Open Service Engine

Open Cloud Platform -4



Conclusions

- INFN e-infrastructure spreads in the entire territory
- Part of an International Collaborative e-Infrastructure
- Open Access & Data "naturally"
- Rich exchange with other disciplines (federation and/or interoperability)
- Capable to study, develop & deploy solutions to demands from Global (Macro) to Local (Micro)