GRDI2020

A Vision for Global Research Data Infrastructures

Interoperability from the e-Science Perspective

Yannis Ioannidis
Univ. Of Athens and ATHENA Research Center

contact@grdi2020.eu www.grdi2020.eu

Outline

- Context:
  - Current situation, vision and future challenges
- Interoperability – definition
- Scientific interoperability / e-Infrastructure layers
- Approach
- Conclusions

Science Paradigms

- 1st - Thousand years ago:
  science was empirical
  describing natural phenomena
  w/ some models, generalizations

- 2nd - Last few hundred years:
  theoretical branch
  using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

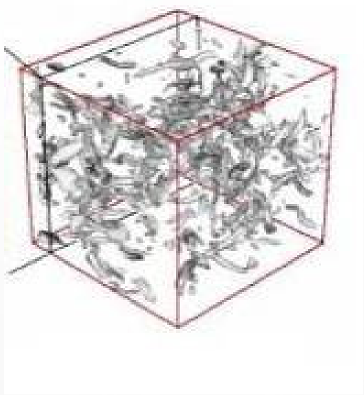
Really Early Times

- One scientist
- One location
- One discipline
- One phenomenon

- One pencil (… carver …)
- One paper (… stone …)
- Street announcements, e.g., Εύρηκα!

Science Paradigms

- 3rd - Last few decades:
  a computational branch
  simulating complex phenomena



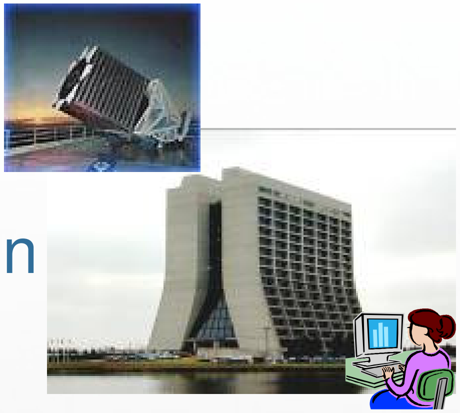
Recent Times

- One small group of scientists
- One location
- One discipline
- One phenomenon

- One file system
- One local disk with custom files
- Publications at refereed forums

Science Paradigms

- 4^{th} - Today:
  data exploration (eScience)
  unify theory, experiment, and simulation



Current Times

- Many/large teams of scientists
- Many locations
- Many disciplines
- Many phenomena

- Many data management systems
- Many data forms
- Web uploads for publications, data, processes, …

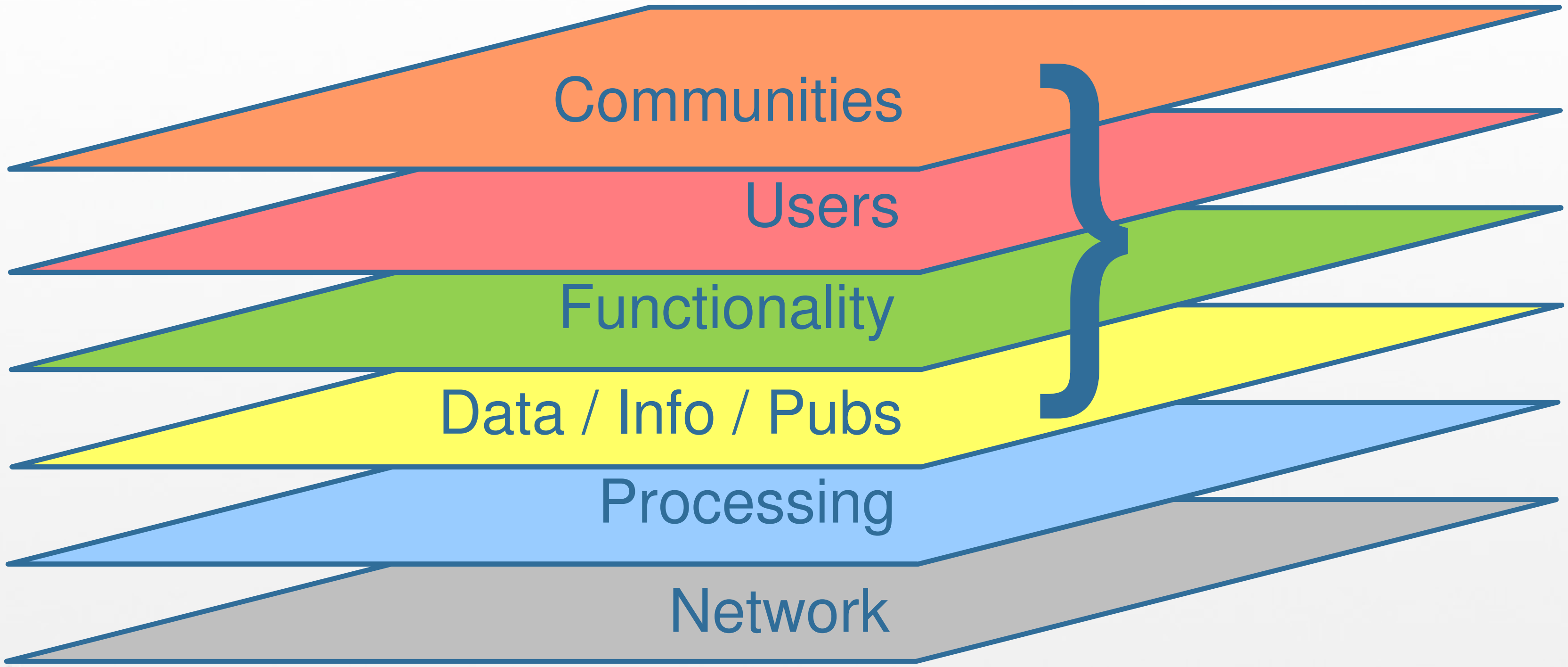
GRDI2020
A Vision for Global Research Data Infrastructures

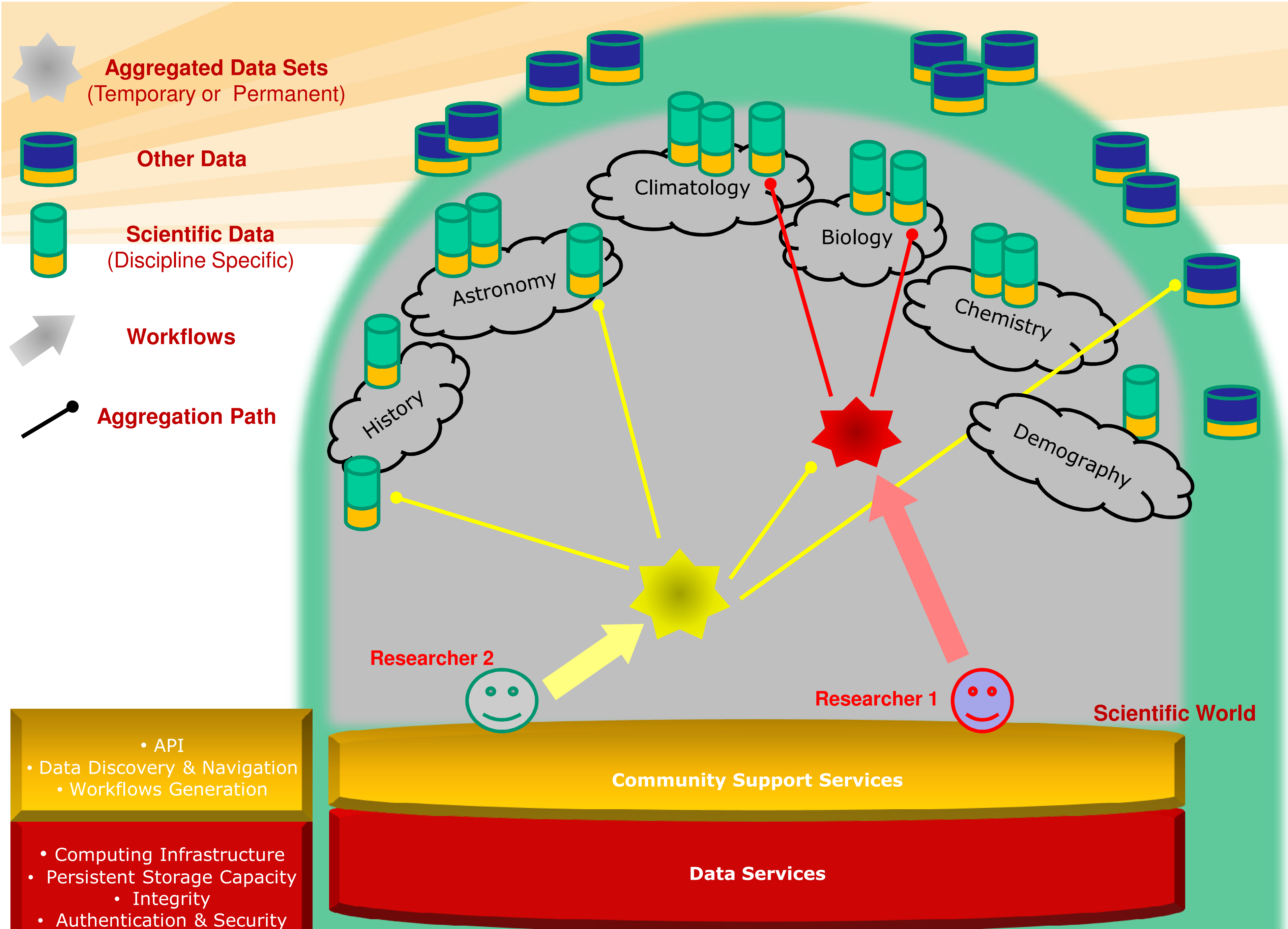
Current Times

- Many/large teams of scientists
- Many locations
- Many disciplines
- Many phenomena
- Many data management systems
- Many data forms
- Web uploads for publications, data, processes, …

- New order in scholarly communication
- Open access
- Creator, author, publisher, curator, preserver roles mixed up
- Digital libraries & repositories at centre stage

eInfrastructure Layers





Source: High-level Group on Scientific Data

Interoperability

DEFINITION

- "*The ability of two or more systems or components to **exchange** information and to **use** the information that has been exchanged*" (IEEE definition)
- *"A property referring to the ability of diverse systems and organizations to work together (inter-operate)" (Wikipedia)*
- "*The capability to communicate, execute programs, or transfer data among various functional units in a manner that requires minimal knowledge of the unique characteristics of those units" (ISO/IEC 2382-2001 Information Technology Vocabulary, Fundamental Terms*)

e-Science

Definition

*"e-Science is computationally-intensive science that is carried out in highly distributed **network** environments, or science that uses immense **data** sets that require **computing**."*

:

- astronomy
- earth science
- biology
- chemistry
- social science
- particle physics
- …

e-Science Interoperability

- "*The ability of two or more e-Infrastructure systems or components to exchange **Scientific** information and to use the information that has been exchanged*"
- Can be applied to different e-Infrastructure "layers"
  - **Network** interoperability
    - Interoperability between different (sub)-networks
  - **Computing** (Grids, Clouds, etc.)
    - Interoperability between different computing systems, e.g. Grids, Supercomputers, Clouds, etc.
  - **Data** interoperability
    - Ability to exchange and use data
- e-Infrastructure service providers need to make interoperability transparent for their users

Network interoperability

- The ability to exchange data between different interconnected networks
  - TCP/IP is the de-facto networking protocol standard enabling world-wide collaboration and data exchange
    - Required a couple of decades
  - However for layers 1 or 2 (end-to-end lightpaths or end-to-end switched connections), interoperability is still not plug-and-play and requires engineering

Computing systems interoperability

- The ability to submit jobs and exchange data between different computing systems such as Grids, Supercomputers, Clouds, etc.
- Indicative areas: Job submission & monitoring, authentication & authorisation, data access
  - No standardised middleware for such components & their interactions: still many years behind for a “TCP-IP”!
  - A variety of architectures, stacks, protocols for different systems
  - There are interoperable implementations, e.g., submitting jobs from one e-Infrastructure to another and vice-versa

Data interoperability

- The ability to exchange data between different systems and use them
- Indicative areas: information representation - profiling, discovery, security
  - Still many challenges ahead
    - Heterogeneity of the exchanged information
    - Inconsistency of the intended use of the information by the two involved systems

Technical Challenges

- **Heterogeneity** of the exchanged information objects
  - Syntactic heterogeneity
  - Structural heterogeneity
  - Semantic heterogeneity
- **Inconsistency of the intended use of the information object**
  - by the producer entity and the intended exploitation of this object by the consumer entity
  - the exchanged information must be complemented with some "**descriptive**" information, such as contextual, provenance, quality, security, privacy, etc. information
  - The descriptive information should be modeled by **purpose-oriented descriptive data models /metadata models.**

Organisational and policy challenges

- Interoperability is not only about technical challenges
- Policy and organisational challenges exist:
  - Interoperability entails (global / con-federated) governance and trust
  - Interoperability entails certification and standardisation
  - Interoperability entails sustained funding and support!

Technical approach: Mediation

- For interoperability a **mediation** concept is needed
  - Includes a mediation function / schema
  - Types of mediation:
    - Mapping
    - Matching
    - Integration
    - Consistency Checking
  - Mediation requires:
    - Adequate modeling of structural, formatting, and encoding constraints of the producer entity information resources
    - Adequate modeling of the consumer entity needs
    - Formally defined transfer and message exchange protocols
    - The definition of a matching relationship between the producer information resources and the consumer models
- **Standards is usually a long and tedious process**

GRDI2020
A Vision for Global Research Data Infrastructures

Conclusions

- E-Science interoperability spans multiple e-Infrastructure layers
- In most of the layers there are still many challenges ahead: above all heterogeneity
- Interoperability is thus complex and never-ending
- For scientific users interoperability should be transparent!

CODATA 22, Cape Town, S. Africa

http://www.codata2010.com/

Thank you!

Questions?