The Data Problem

A View on Applying AA(A) Technologies to Data Infrastructures

Telefónica

Telefónica I+D
12.07.2012

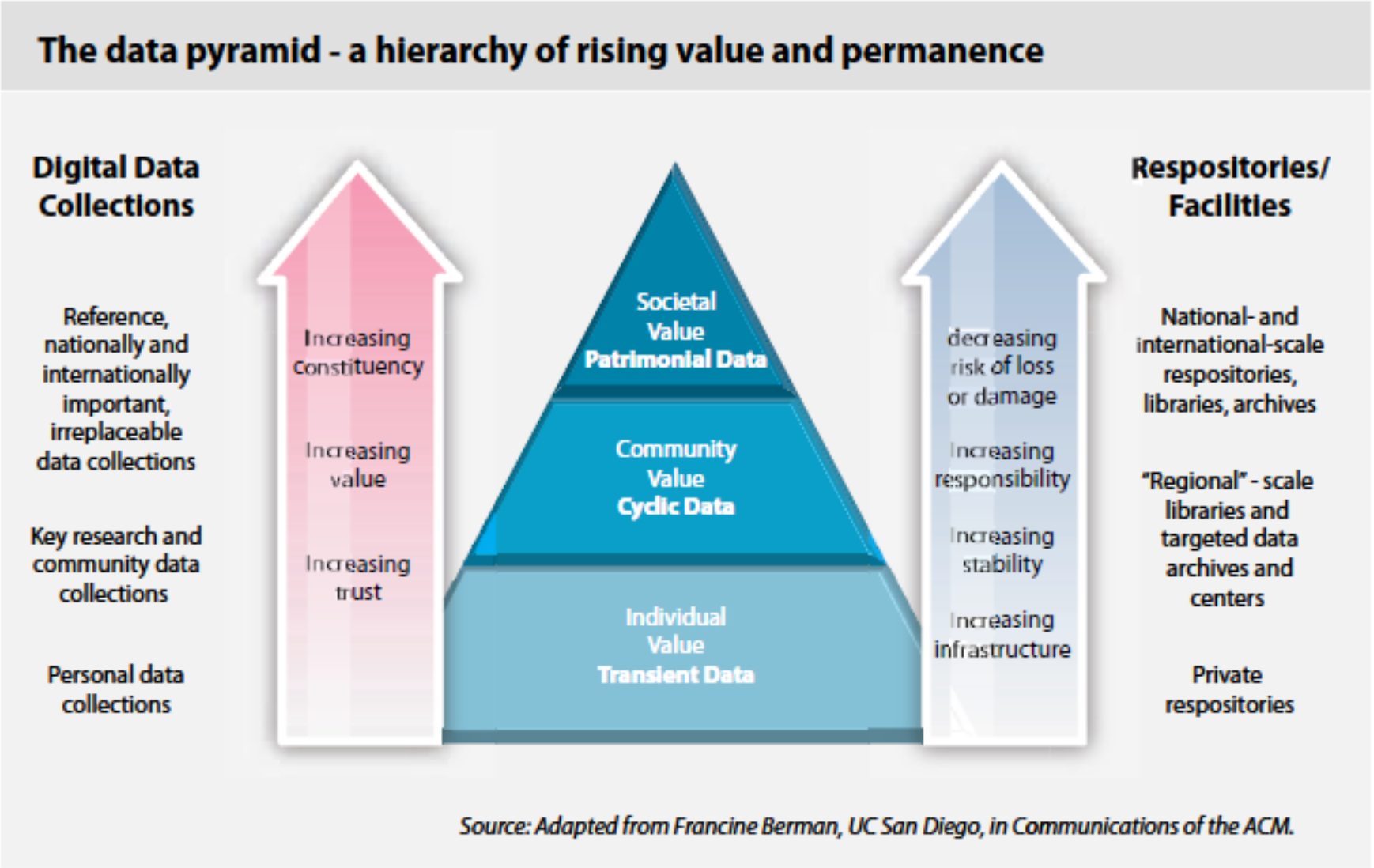
The Long-Term Vision

"Our vision is a scientific e-Infrastructure that supports seamless access, use, re-use and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance."

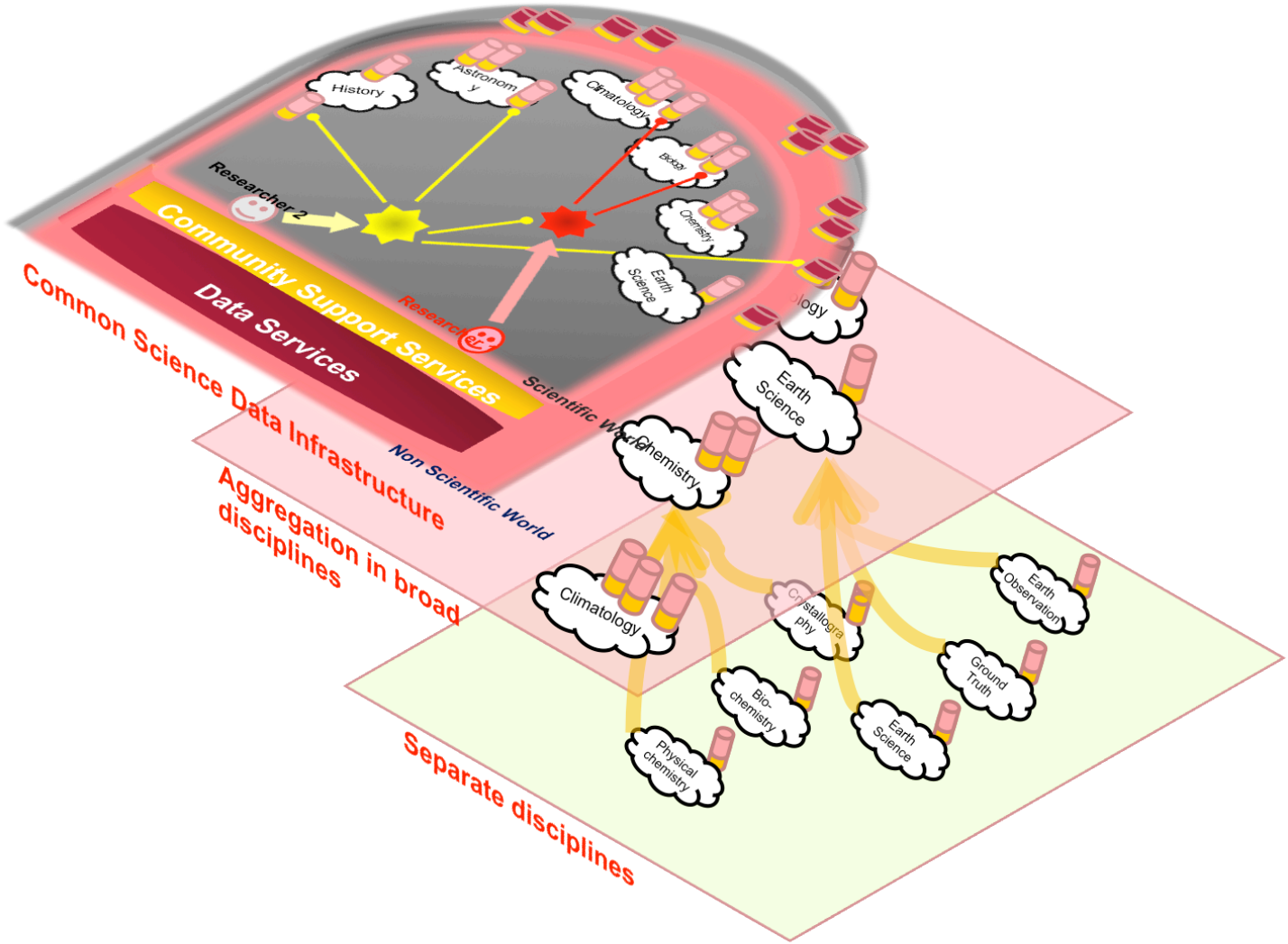
Report of the High-Level Group on Scientific Data, October 2010

"Riding the Wave: how Europe can gain from the raising tide of scientific data"

The Data Pyramid



Aggregating and Integrating



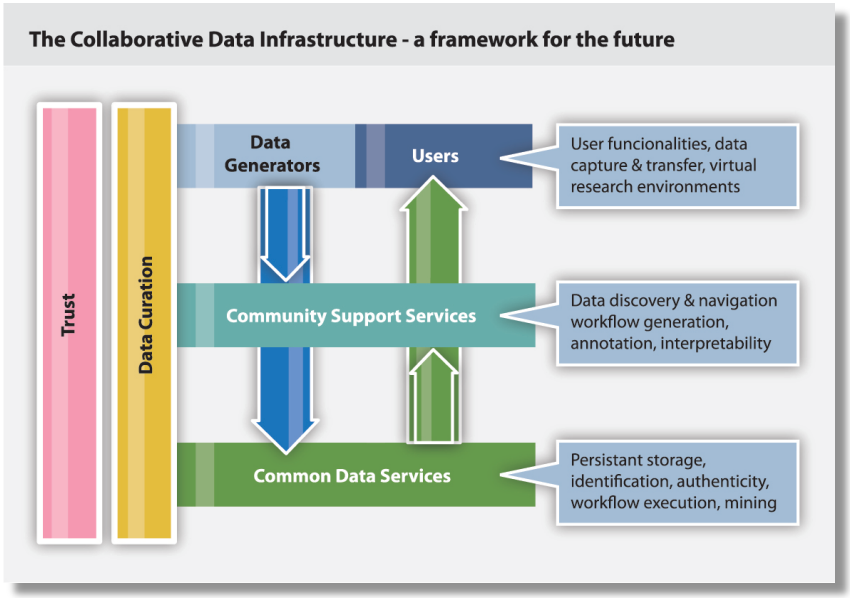
An Initial Wish List

- Open deposit, allowing user-community centres to store data easily
- Bit-stream preservation, ensuring that data authenticity will be guaranteed for a specified number of years
- Format and content migration, executing CPU-intensive transformations on large data sets at the command of the communities
- Persistent identification, allowing data centres to register a huge amount of markers to track the origins and characteristics of the information
- Metadata support to allow effective management, use and understanding
- Maintaining proper access rights as the basis of all trust
- A variety of access and curation services that will vary between scientific disciplines and over time
- Execution services that allow a large group of researchers to operate on the stored date
- High reliability, so researchers can count on its availability
- Regular quality assessment to ensure adherence to all agreements
- Distributed and collaborative authentication, authorisation and accounting"
- A high degree of interoperability at format and semantic level

Adapted from the PARADE White Paper

Identity in Data Infrastructures

- The “Riding the Wave” report includes eleven challenges to overcome in its vision of a Scientific Data Infrastructure
  - From data collection to sustainability
- Three of them directly rely on the availability of an operational identity infrastructure
  - Trust
  - Security
  - Data publication and access
- At least other four would greatly benefit from it
  - Usability, diversity, new social paradigms, and preservation and sustainability



Data and AA Infrastructures

- Security, trust and access requirements must become pervasive and dynamically associated with data themselves and their metadata,
  - Entities can apply the policies they consider relevant
  - Derived datasets generate their security metadata from the origin dataset(s)
  - Coherent availability of information across different access patterns
- The accuracy of these metadata requires well-established identity services
  - Not only people
  - Not only through browsers
  - Data sources
  - Data processors
  - Datasets themselves
- An Authentication and Authorization Infrastructure (AAI) becomes a key element for the Data Infrastructure
  - At all layers in the data pyramid
  - In all levels of aggregation and integration

The Ideal State of Several Well-Known Parts

- Entities prove their identities at identity providers (**IdPs**)
  - By different means
  - Security and usability criteria.
- Identity attribute values are stated by attribute authorities (**AAs**)
  - Under control of different organisations and communities
  - Communities should be empowered to manage their AAs as they see fit
- Identity is transferred to applications by relying parties (**RPs**)
  - They accept data from recognized IdPs and AAs
- RPs and applications rely on Policy Decision Points (**PDPs**)
  - Managed by rules defined by data infrastructure operators and dataset owners
  - PDP rules are manipulated according to human-friendly mechanisms
- A trust framework must be established among the interacting parties
  - The foundation of all identity data exchange and application
  - A *heterarchical* model is the only possible one

Don’t Forget the Third A (and beyond)
Accounting and Reputation

- Accounting is an essential function to keep
  - Audit capabilities: Funding, resource location…
  - Security: Incident handling, forensic evidence…
  - Usage records: Resource planning, access methods…
  - Resiliency: Diagnostics, fault isolation….
- Accounting logs are datasets themselves
  - Benefit from data infrastructure tools
  - Complex querying, aggregation and integration
  - And access control
- Reputation expresses the value a certain community gives to an entity
- Reputation shall come from a web of trust built of several sources
  - Formal: Peer reviews, project reports…
  - Informal: Social tags, user ratings…
  - Infrastructural: Availability, incident reports…
- Reputation becomes, in essence, a special type of accounting record
  - Similar characteristics
  - And a relevant attribute source

Keep It Simple. Let It Happen

- Avoid technical and administrative overheads
  - Do not reinvent the wheel in several shapes
  - Take very careful steps into additional formalizations
  - Experience shows how overcomplicated proposals have been sidestepped
- Take advantage of existing trust links in the scientific community
  - And of its specific characteristics
  - Build on strong foundations
- Take the outer worlds into account
  - Governmental infrastructures
  - Commercial initiatives
  - Other environments: professional associations, NGOs…
- Make the whole system sustainable
  - Sustainability is by definition opposite to perpetual growth

A Few Immediate Challenges

- Expanded user authentication
  - edu + (com + gov)
  - Multiple sources of identity data
- Applicable authorization
  - Policy engines available to users and communities
  - Interpretable rules
- Reliable accounting
  - Common operational baseline
  - Accounting data as metadata
- Reputation support
  - Formal and informal records
- For humans and non-humans
  - Any entity needs an identity
- Transparent integration with any other infrastructure
  - Network, computational,…
  - Make composition simple
- Keep It Happening