RIDING THE WAVE

AFTER THE EU HLEG REPORT
VISION (AND REALITY) ABOUT ACCESSING RESEARCH DATA

Peter Wittenburg, (Krister Linden)
The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Outline

- **Group and Motivation**
- The Research Data World
- Opportunities and Challenges
- Collaborative Data Infrastructure
- Other Relevant Aspects
- Vision 2030
- Action Points

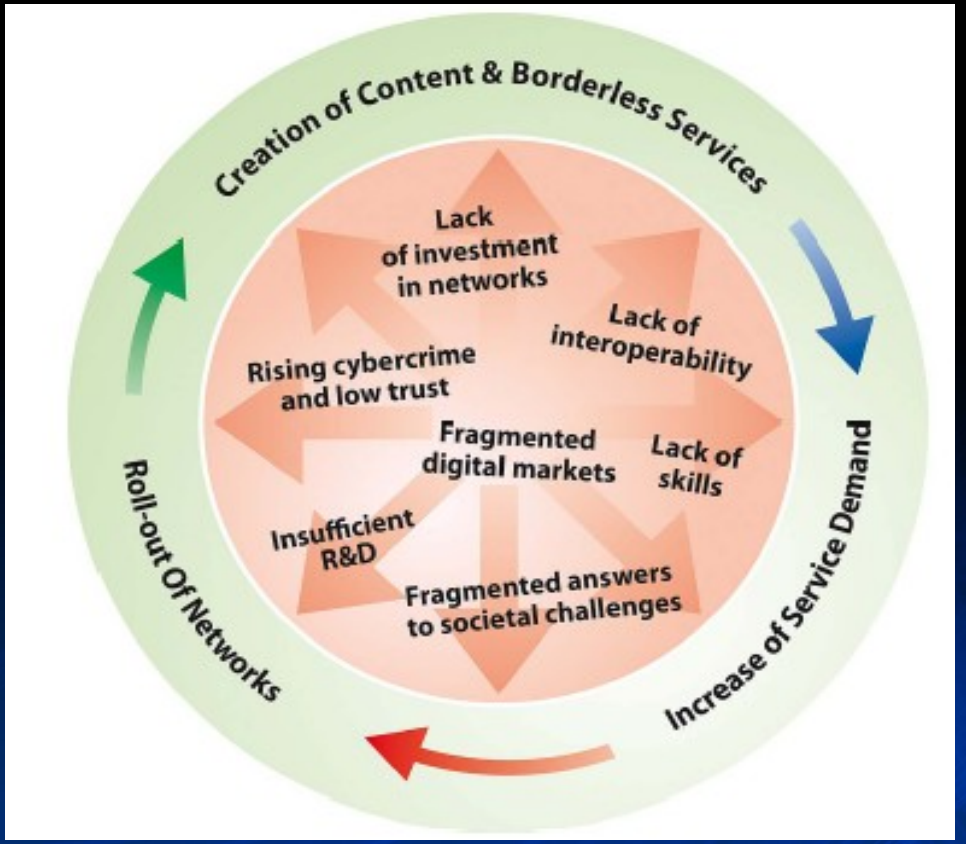
HLEG and Motivation

- EC DG InfSo invited 11 experts and 12 contributors to 6 meetings
  Chair: John Wood; Rapporteur: David Giaretta; Members: Thomas Andersson ; Achim Bachem; Christoph Best ; Françoise Genova ; Diego R. Lopez; Wouter Los; Monica Marinucci; Laurent Romary; Herbert Van de Sompel; Jens Vigen; Peter Wittenburg; DG InfSo Representatives: Konstantinos Glinos; Carlos Morais-Pires;

- Goals
  - come up with a vision 2030 for the management of research data as a guideline for future actions of the EC
  - discuss all relevant aspects around “data” in an unbiased manner
  - accelerate measures to take care of our data and to remain competitiv
- Motivation
  - enormous increase in scale and complexity
  - not only summarize what some of us already know or are doing, but facilitate a systematic and global approach and push ahead actions
  - knowledge is power - data has a value although difficult to quantify

Digital Agenda for Europe
the policy context

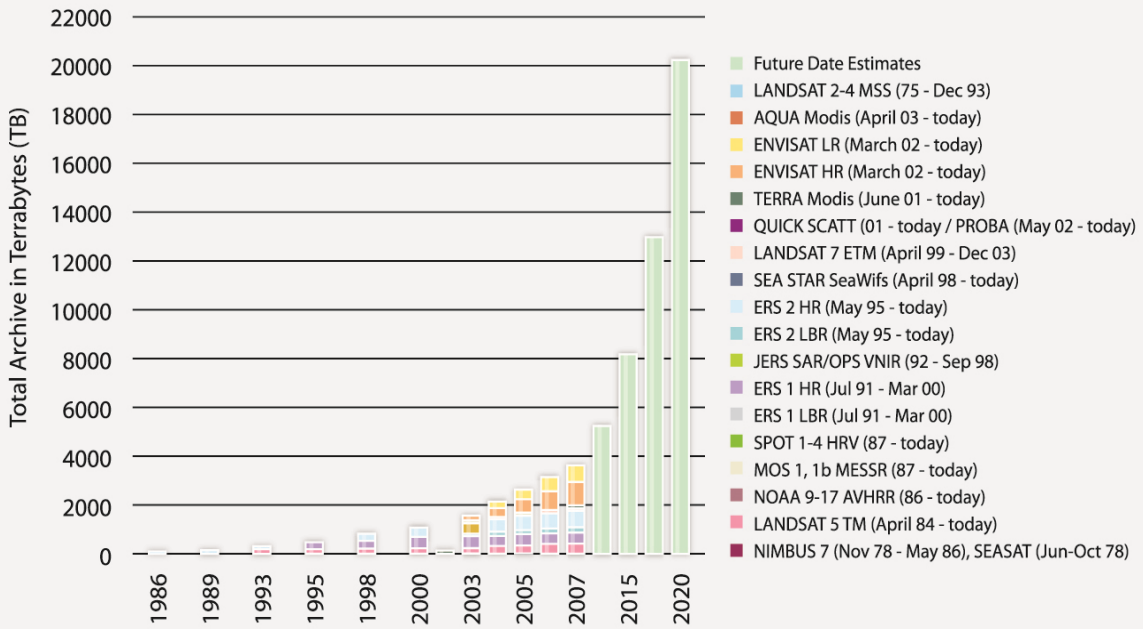
DAE is one of the flagships of **"Europe 2020: a strategy for smart, sustainable and inclusive growth"**



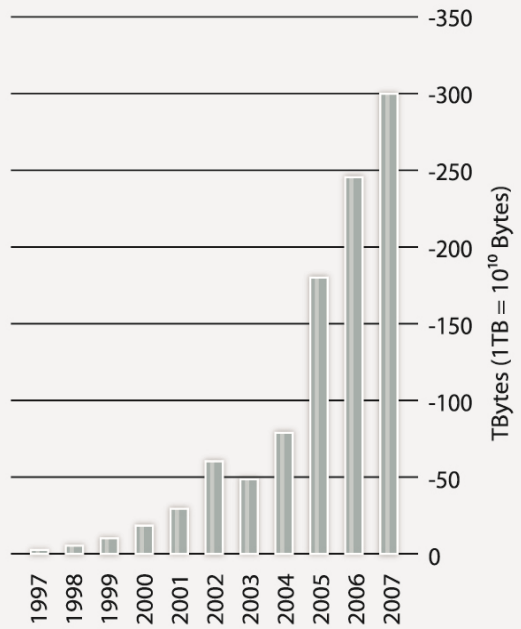
Trends are Known

"A fundamental characteristic of our age is the raising tide of data – global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge."

Evolution of ESA's EO Data Archives between 1986-2007 and future estimates (up to 2020)



Yearly Data Creation on NICE

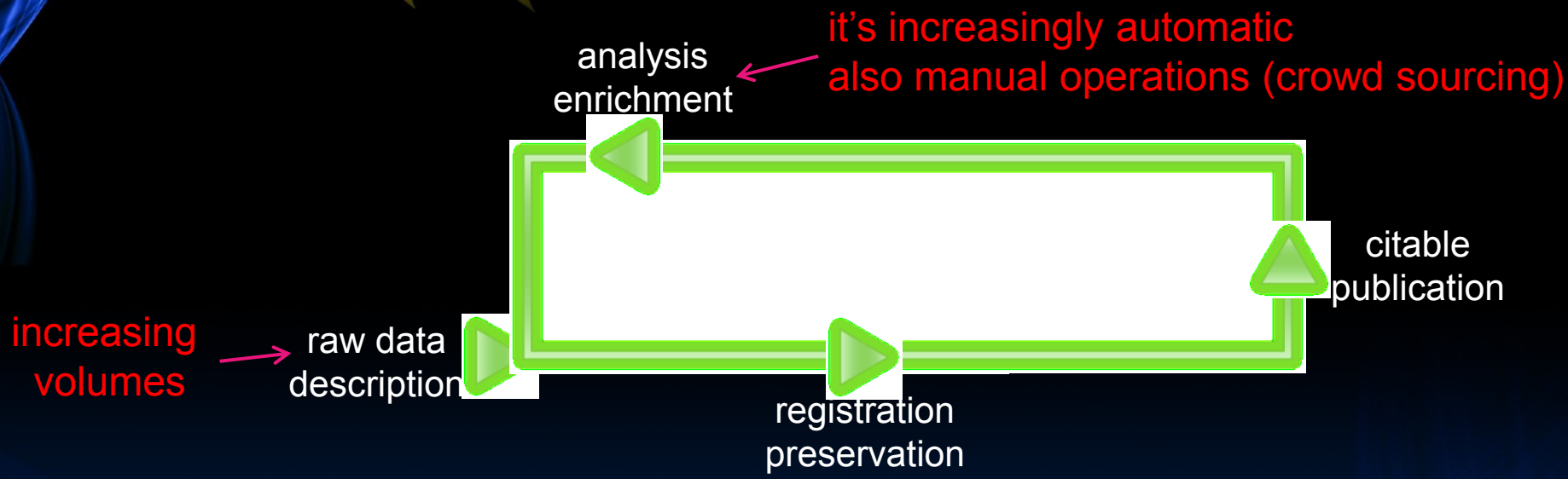


Outline

- Group and Motivation
- **The Research Data World**
- Opportunities and Challenges
- Collaborative Data Infrastructure
- Other Relevant Aspects
- Vision 2030
- Action Points

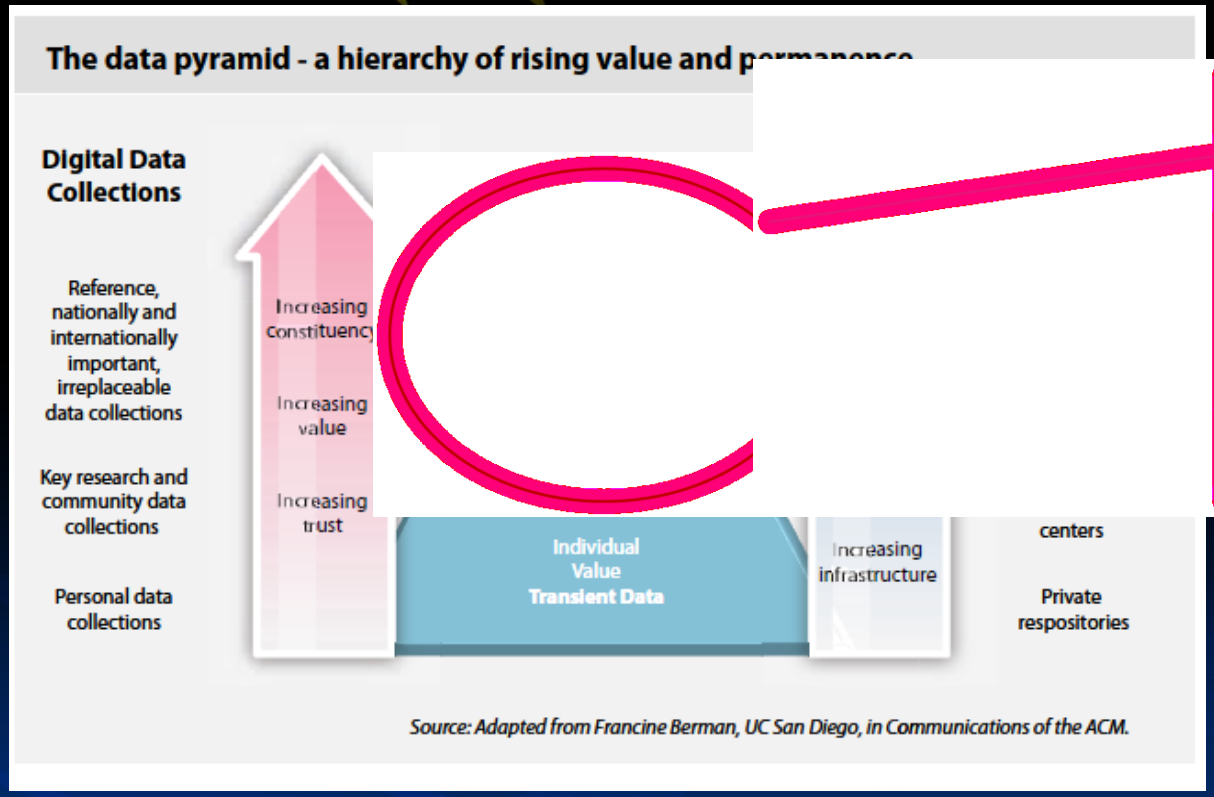
Research Data World

- Knowledge Cycle is changing



- Exabyte scale and millions of related files of different types create unseen complexity - deal with a new quality
- much relevant data is and will not be registered
  (80 % of recordings about languages and cultures are endangered)

Berman's classification



this is the data we need to take care of
but do we know which data will be of relevance for future generations?

Research Data World

- some interesting aspects
  - lossless separation of content and carrier in the digital domain changes the world - some speak about a revolution comparable with the invention of book printing
  - data creators are not known personally to data users anymore - we ne to solve the trust problem
  - there is no doubt: data accessibility changes nature, pace and directio of research
  - diversity in many dimensions is the dominant feature of scientific information and this will probably increase due to the inherent innovation forces
  - technology allows to include the citizens in different roles - also as contributors, increasing volume and complexity
  - increasing pressure towards open access

Outline

- Group and Motivation
- The Research Data World
- **Opportunities and Challenges**
- Collaborative Data Infrastructure
- Other Relevant Aspects
- Vision 2030
- Action Points

Opportunities and Challenges

- virtual integration
  - integrating large data sets across disciplines and countries to create new insights
  - recombining data to virtual collections from different perspectives
  - sufficient data as basis for holistic modeling and understanding
  - data intensive science: find correlations and draw inferences not constraint by pre-assumptions - huge amounts of data not used
- tackling the grand challenges resulting from human activities
  - climate change, sustainable energy, stability health, etc.
  - stability of our societies and minds given the innovation, changes, globalization and migration
- facilitating the many "small research questions" driven by scientific curiosity
- relieve researchers from data management and curation effor

Opportunities and Challenges

- however there are quite some hurdles to overcome
  - need to change culture and researchers minds to deposit data
  - need to establish trust at depositor's and user's side
  - trust has to do with data quality, integrity and authenticity
  - need to convey context and provenance to allow users to understand
  - need new responsibilities and new mechanisms to solve
    data curation, preservation, organization and granting access
    without ignoring security and ownership principles

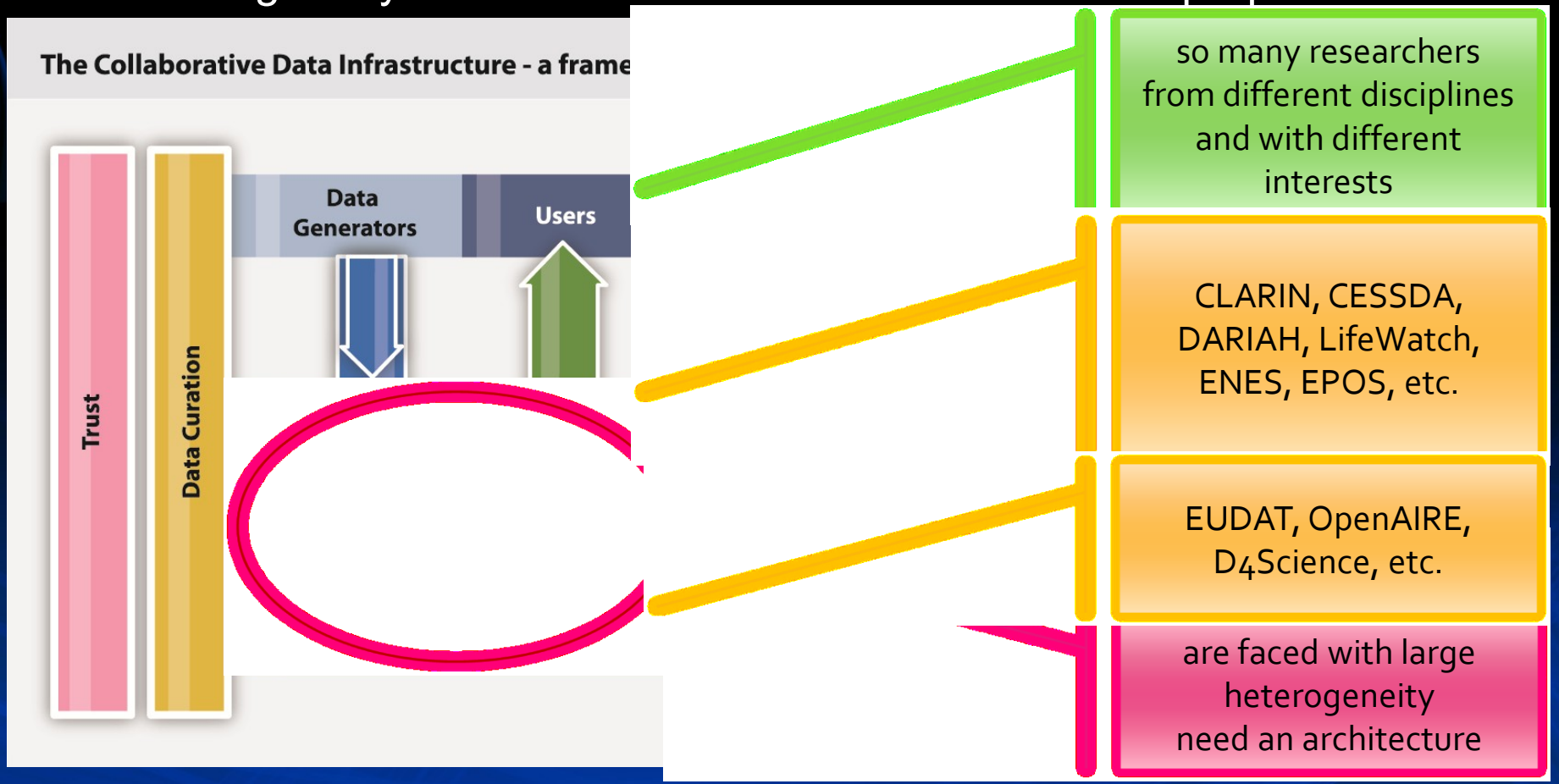
  - need incentives for researchers to deposit in proper quality so that
    data publication helps in career and reputation building

Outline

- Group and Motivation
- The Research Data World
- Opportunities and Challenges
- **Collaborative Data Infrastructure**
- Other Relevant Aspects
- Vision 2030
- Action Points

Collaborative Data Infrastructure

- obviously we need a new layer of responsibility:
  a systematically constructed and global data infrastructure
- some already working on data organizations - piecemeal, fragmented
- we call it a Collaborative Data Infrastructure open for many players and heterogeneity based on an abstract architecture and proper APIs



Collaborative Data Infrastructure

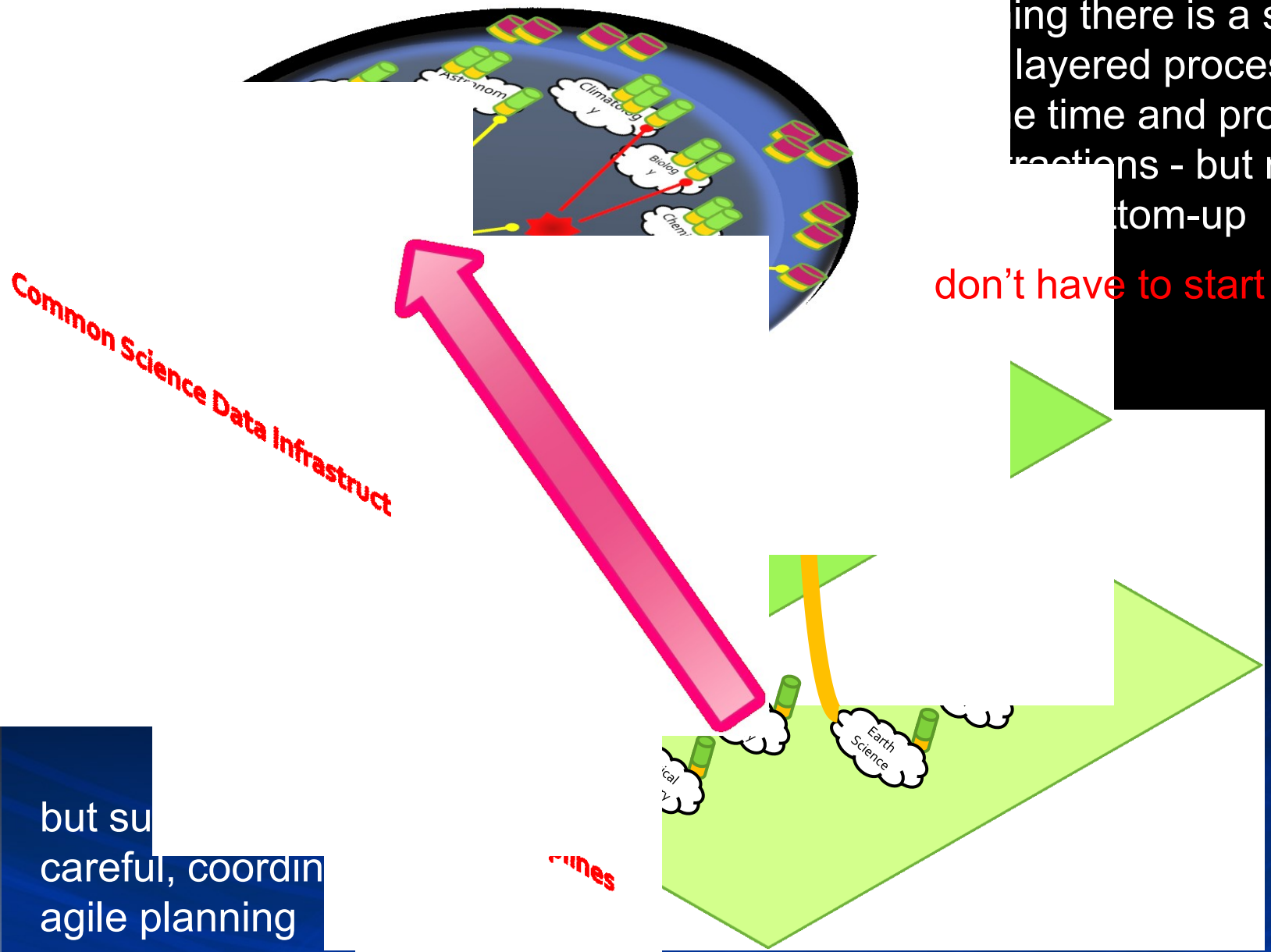
- some requirements/characteristics
  - balance flexibility and reliability, secureness and openness, local operation and global integration, affordability and high performance
  - integration and interoperability require standards and agreements
    nevertheless CDI may not hinder flexibility and openness
  - CDI needs to preserve data, guarantee authenticity, detect unauthorized alterations, give access, define online levels of access, help in understanding, protect privacy
  - CDI needs to be distributed and offer redundancy
  - community layer to be cross-discipline and cross-border
  - there is no single technology that fits all - but fragmentation is inimical
  - distributed based on organic processes - not one central organization
  - need to define interoperability:
    can only be at data object and not at data content level

let's not look in detail
will not be trivial to achieve

Collaborative Data Infrastructure

ing there is a step-wise
layered process requiring
e time and proper
ractions - but needs to be
ttom-up

don't have to start from scratch



but su
careful, coordin
agile planning

Outline

- Group and Motivation
- The Research Data World
- Opportunities and Challenges
- Collaborative Data Infrastructure
- **Other Relevant Aspects**
- Vision 2030
- Action Points

Other Relevant Aspects

- funding
  - need to understand data as a socio-economic treasure in a competitiv domain - at the end research is about global competition
  - need proper business models - who is paying, which data is free, etc.
  - governments will have to reserve funds for data management
- quality and impact
  - need to measure quality and impact, which metrics are meaningful
  - need to reward contributors
- management/curation skills
  - need a new type of experts: data scientists
- power researchers
  - resulting CDI will be complex as the data world will be
  - need to educate and train a new generation of power users
- ecology
  - uncontrolled copying of data sets is not ecological
  - need to take care of green computing principles

Outline

- Group and Motivation
- The Research Data World
- Opportunities and Challenges
- Collaborative Data Infrastructure
- Other Relevant Aspects
- **Vision 2030**
- Action Points

Vision 2030

All stakeholders, from scientists to national authorities to general public are aware of the critical importance of preserving and sharing reliable data produced during the scientific process.

Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data and they can evaluate the degree to which the data can be trusted.

Producers of data benefit from opening it to broad access and prefer to deposit their data with confidence in reliable repositories. A framework of repositories work to international standards, to ensure they are trustworthy.

Vision 2030

Public funding rises, because funding bodies have confidence that their investments in research are paying back extra dividends to society, through increased use and re-use of data.

The innovative power of industry and enterprise is harnessed by clear and efficient arrangements for exchange of data.

The public has access and can make creative use of the huge amount of data available; it can also contribute to the data store and enrich it.

Policy makers can make decisions based on solid evidence, and can monitor the impacts of these decisions.

Global governance promotes international trust and interoperability.

Outline

- Group and Motivation
- The Research Data World
- Opportunities and Challenges
- Collaborative Data Infrastructure
- Other Relevant Aspects
- Vision 2030
- **Action Points**

Action Points

- HLEG requests
  - need a CDI
  - earmark additional funds
  - develop new ways to measure data value and reward researchers
  - train a new generation of data scientists and broaden understanding
  - think green
  - establish a high level coordination group

- Recent Developments
  - EUDAT got funds to work on CDI (15 communities, 8 centers)
  - work on establishing a Data Access and Interoperability Task Force
    - DAITF part of a rich landscape (IETF, ITU, OAI, ISO/IEC, etc.)
    - example: ITU has signed an agreement to collaborate on a Digital Object Architecture

Thanks for your attention.