CESGO : UN ENVIRONNEMENT VIRTUEL DE RECHERCHE POUR LES SCIENCES DE LA VIE

CargoDay Rennes

19/11/2015 - 11h10/11h40

Olivier Collin / Yvan Le Bras - Plate-forme Bioinformatique GenOuest
Olivier.Collin@irisa.fr

Context

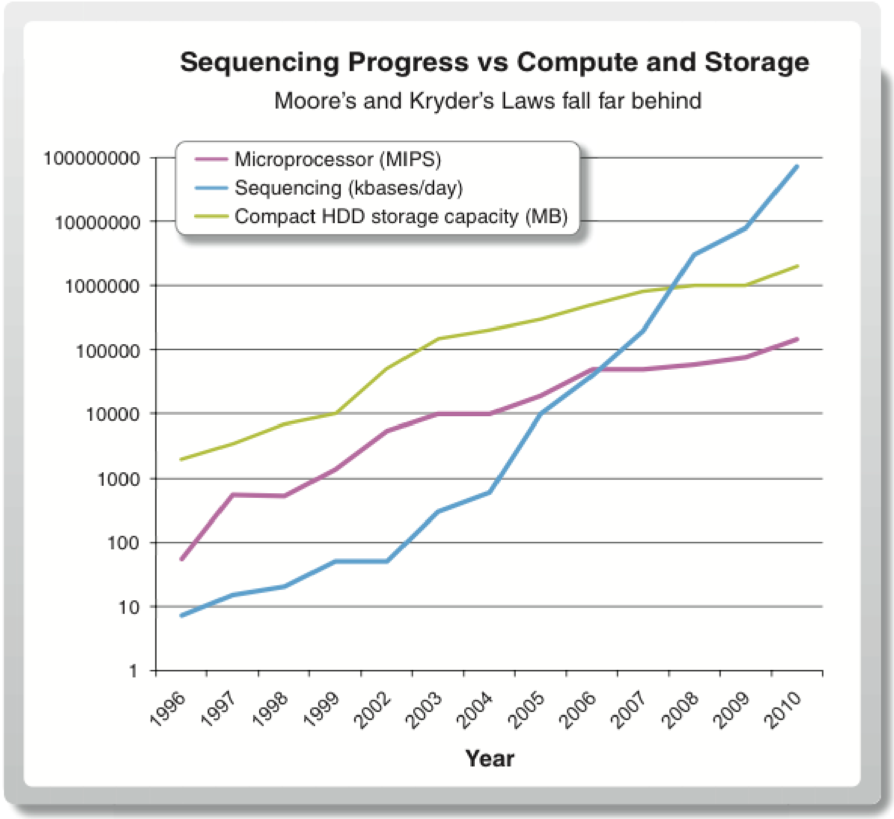


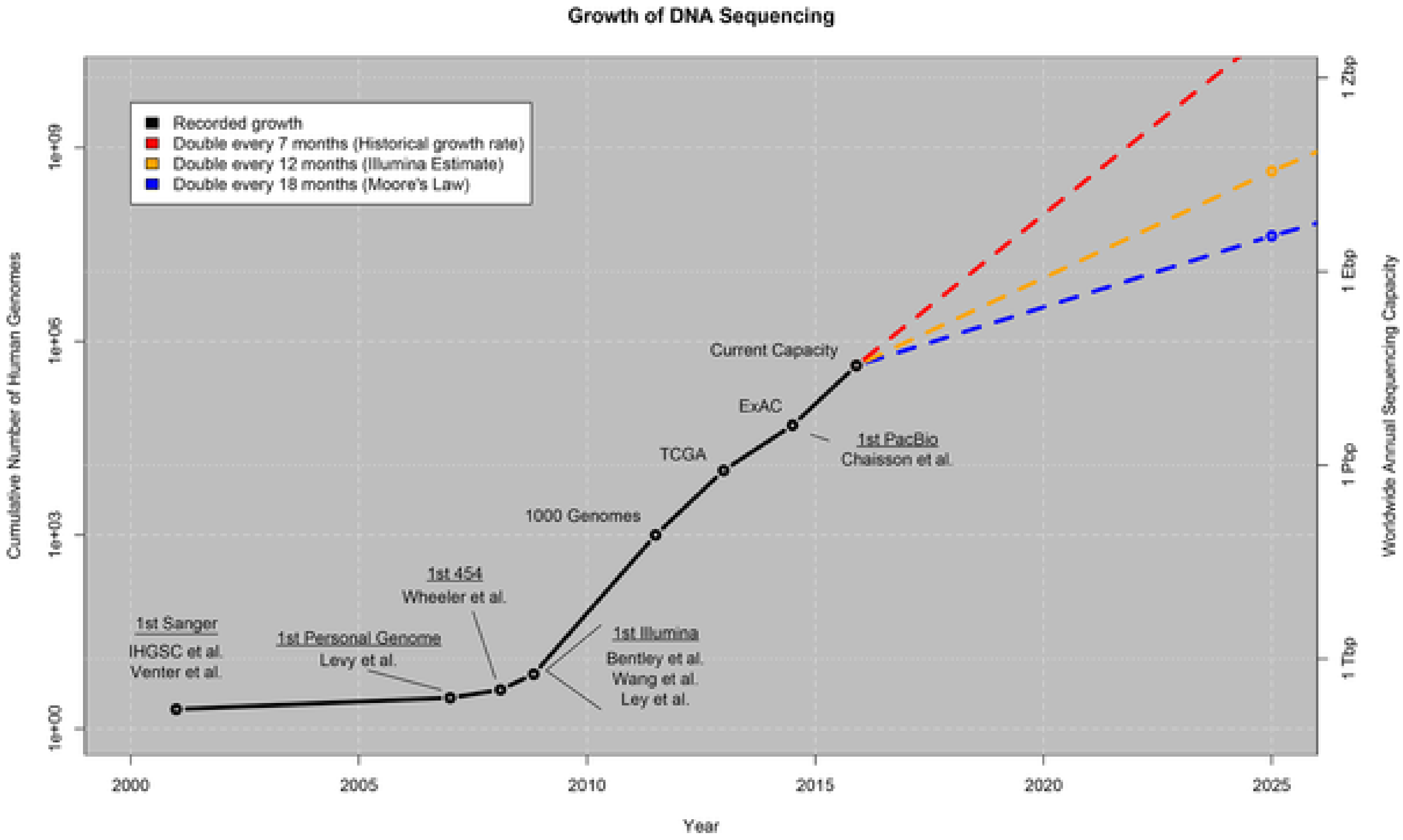
Fig. 1. A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

Kahn. On the future of genomic data. Science (2011) vol. 331 (6018) pp. 728-9

- Now : Genomics : Next Generation Sequencing
- Next : Proteomics
- Next : Bio-imaging

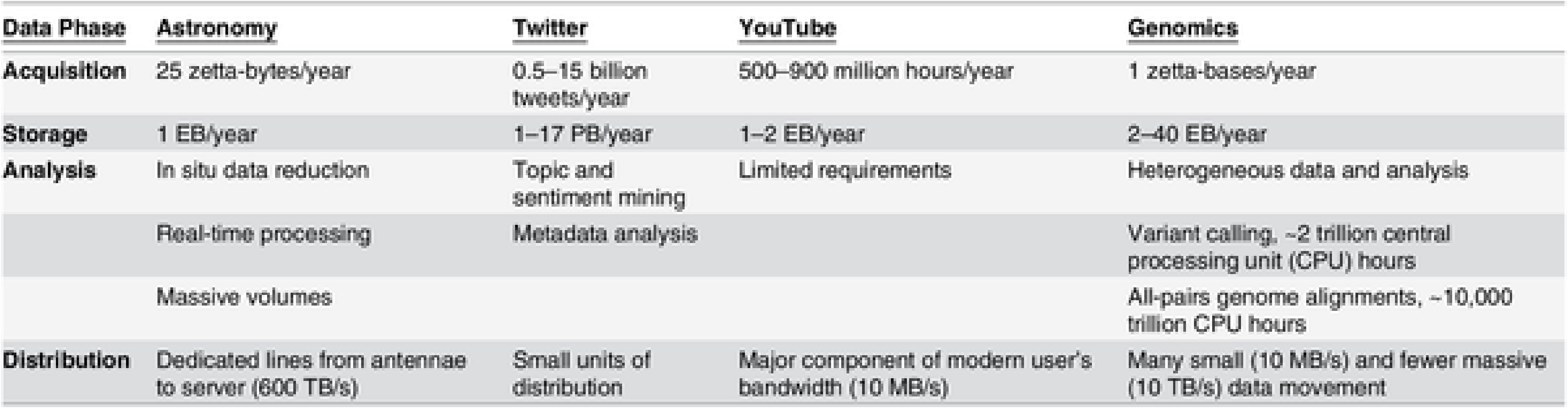
- Digital data
  - Huge amount
  - Heterogenous
- Critical situation for some laboratories

Fig 1. Growth of DNA sequencing.



Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195
http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195

Table 1. Four domains of Big Data in 2025.



Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195
http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195

Challenges

Biology becomes a digital science

- New technologies with lower costs and high throughput create both a formidable scientific opportunity and a dangerous situation.

Evolution

- Automatization implies more human resources for data analysis
- Need of technical competences often missing in Biology research laboratories
- Evolution of the biologists work
- New skills and competences

How to switch from a discipline structured for data production to a discipline structured for data analysis ?

E-Biogenouest/CeSGO

- Project started in May 2012 for 3 years
- Funded by Brittany and Pays de la Loire
- E-science initiative for the Biogenouest network
- Bottom-up approach

- Roadmap preparation
- Community building
- Training/workshops
- Experimentation/Pilot project : Virtual Research Environment (VRE)
- CPER funded by Brittany Region, INRIA, Europe.

VRE : definition

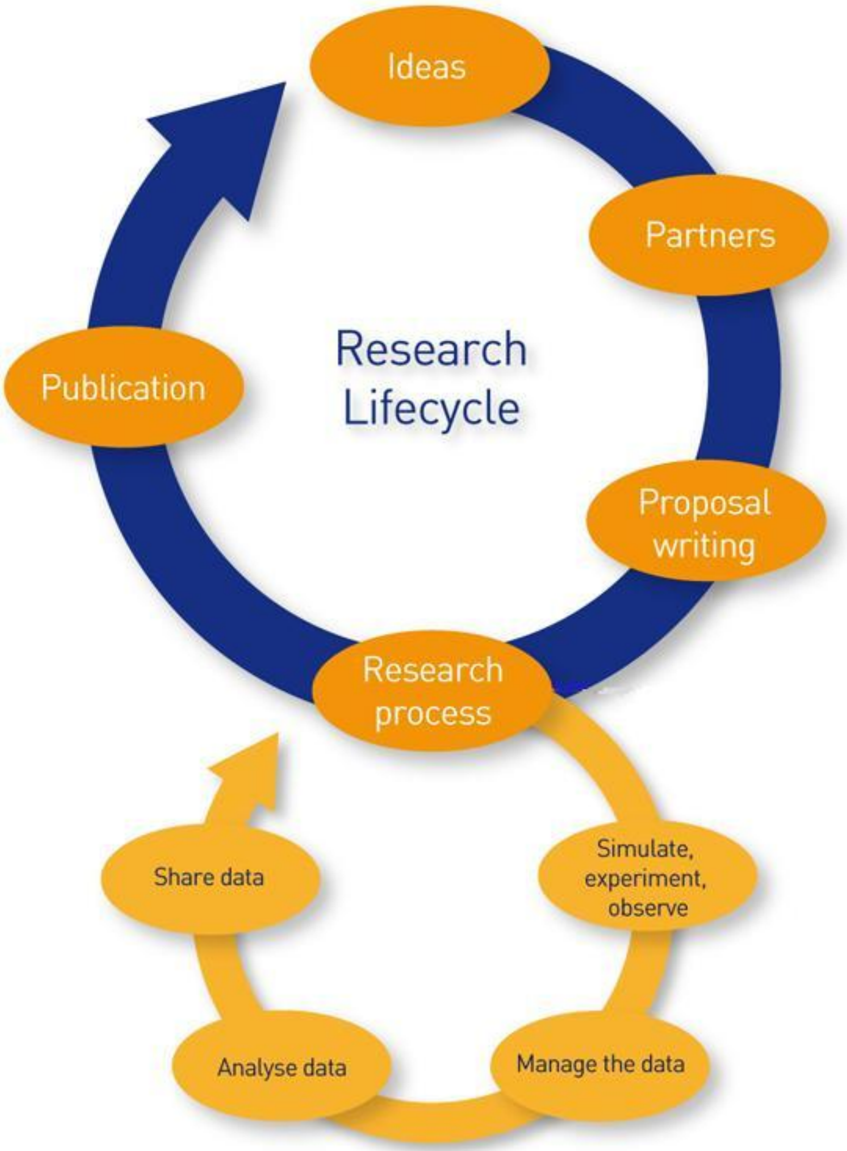
Candela et al. Virtual Research Environments: An Overview and a Research Agenda. Data Science Journal 01/2013; 12:GRDI75-GRDI81

Collaboration tool for scientists

- Web based
- Support communities of practice
- Resources adapted to the communities needs
- Open and flexible
- Support fine-grained controlled sharing of resources

VRE

Strong interactions with the research life cycle

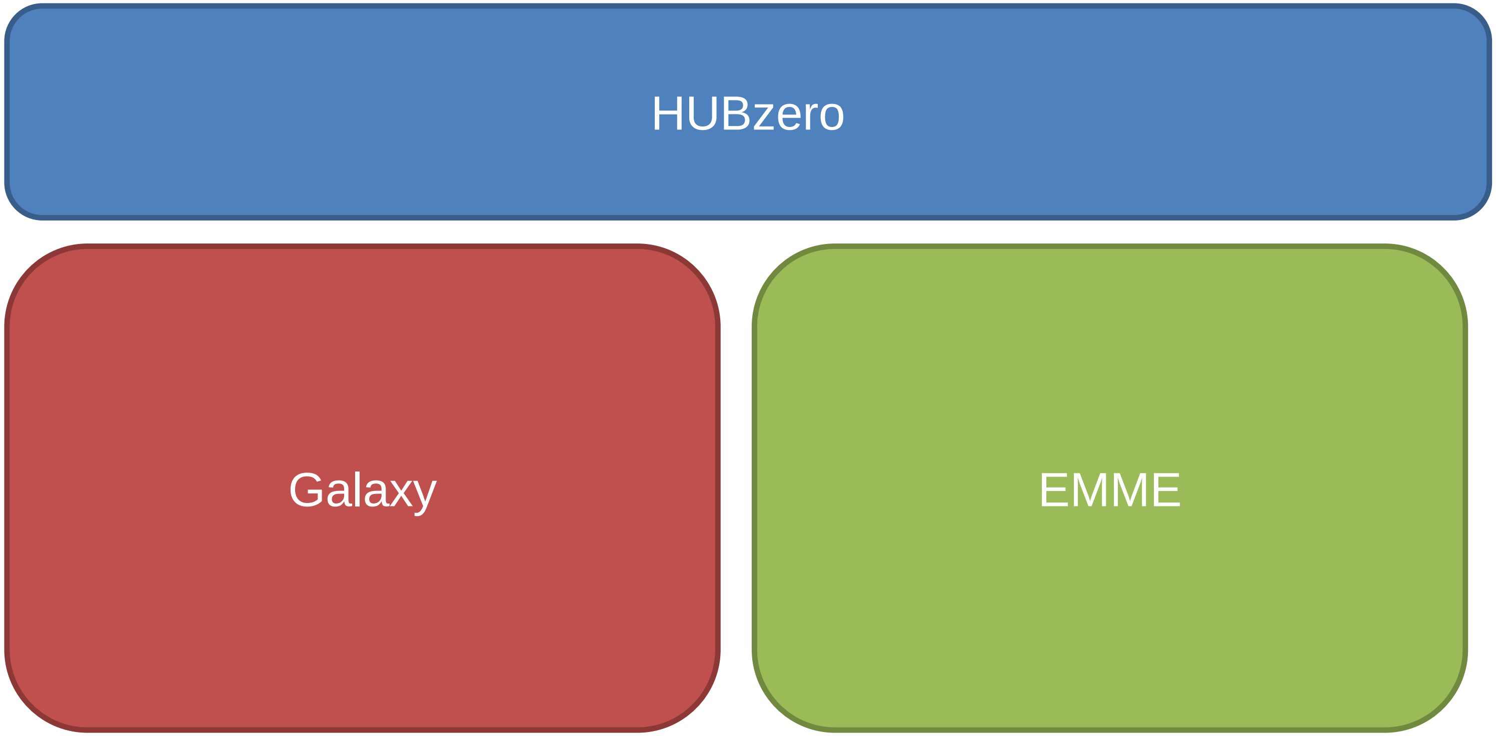


http://tools.jiscinfonet.ac.uk/vre-lifecycle/index.html

A system of systems

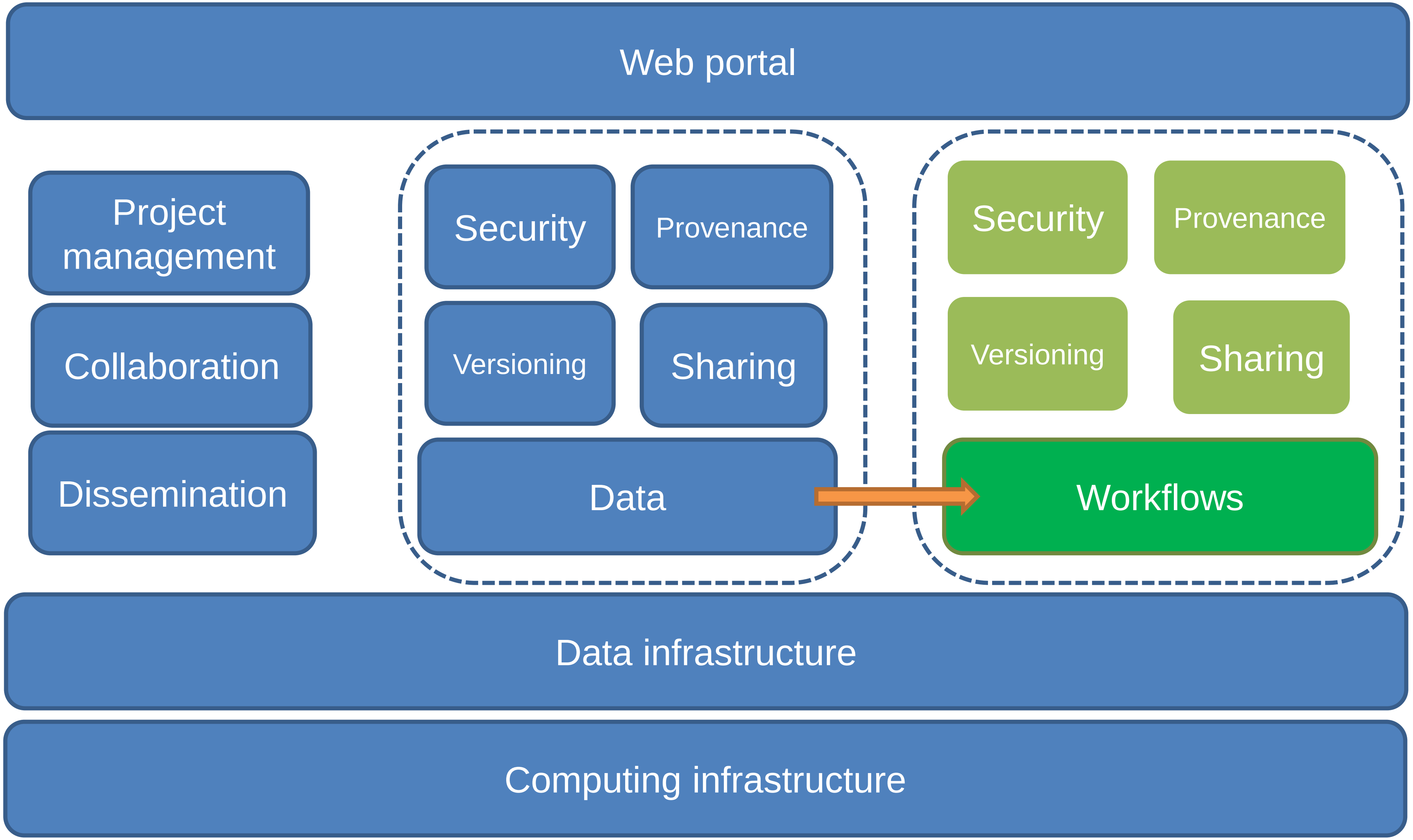
- Combination of various tools
  - A data analysis portal : Galaxy
  - A metadata management tool : ISAtools suite
  - A collaborative portal : HUBzero
  - Additional utilities :
    - Pydio : file transfer
  - Some software glue to make it work…
    - BioBlend : Galaxy API
    - In-house developments

Continuum



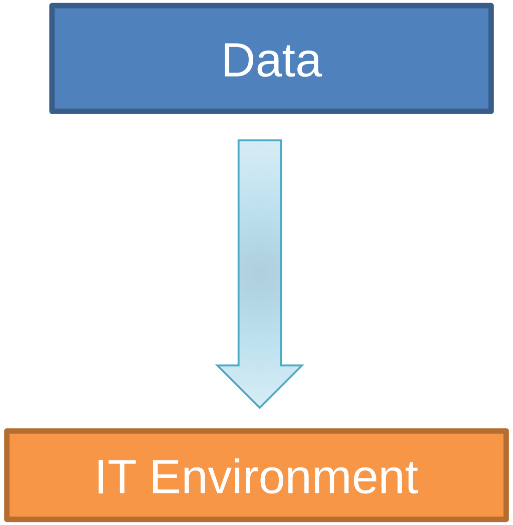
- Continuum for the management and analysis of biological data
- Collaborative environment

VRE : Virtual Research Environment



A paradigm shift

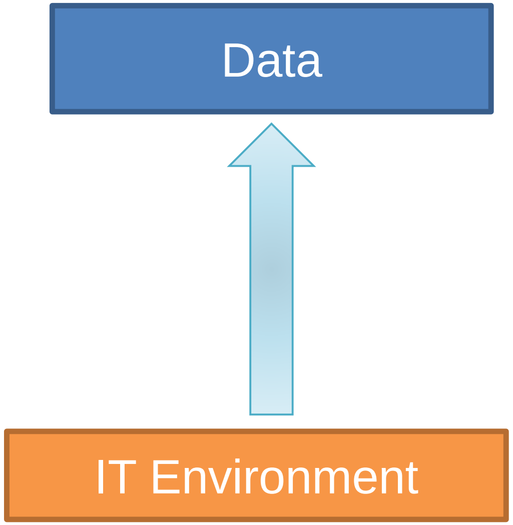
From…



Data

IT Environment

To…



Data

IT Environment

« Bringing back Biology to the biologist »

New paradigm

- New characteristics of Biology
  - Data intensive science
  - Need for multidisciplinary interactions and increased collaboration
  - Sharing and openness
- Data
  - Should become a « *first class citizen* » of scientific communication
  - Should be discoverable : DOI : digital objects identifiers
- But…

  Which incentives exists to make researchers freely share their data ?

E-Biogenouest VRE and Open Science

VRE (HUBzero + ISAtools + Galaxy)

- Everything is a shareable resource for HUBzero
- HUBzero supports DOI
- Metadata in ISAtools supports provenance
- Computational workflows are shareable in Galaxy

- Scitizen : citizen science, a science collaborative portal
  - a scientist can create free projects, and build a community of users to collect data
  - Definition of a form, user sends a picture (position and the filled form
  - Target domains : biology, ornithology, botany, architecture, archeology, etc.

Current situation

- VRE for Life Sciences
  - 200 users / 800 resources
  - From e-biogenouest.org to CeSGO
  - https://www.e-biogenouest.org/
  - http://cesgo.genouest.org/
- CPER :
  - Equipment
  - Lack of human resources
- VRE workgroup of IFB (Institut Français de Bioinformatique)
- Steering committee of UEB VRE project (Appels Innovants) : 8 months project

Future

- E-science facility
  - Focused on data management and data analysis
    - Findable, Accessible, Interoperable and Reusable (FAIR)
    - Data Management Plans implementation (contacts DCC)
    - DOI attribution (contacts INIST)
    - Trusted Data repository (forget it…)
  - Multidisciplinary interactions
    - Digital science needs expertise
  - Open to society
    - Citizen science

Example 1 : Scitizen : framework for Open Science

Example 2 : From MMORPG (Massively MultiPlayer Online Role Playing Game) to MMOS (Massively MultiPlayer Online Science)

http://mmos.ch/

Goals

- For society
  - Open Science & Open Data
- For end users scientists communities
  - Data management plan
  - Preserve, access, share & visualize (data and analytics processes)
  - Project management
- For ICT
  - Ease the use of tools
  - Accelerate the switch from dev to production
  - Optimize the infrastructure

Conclusion

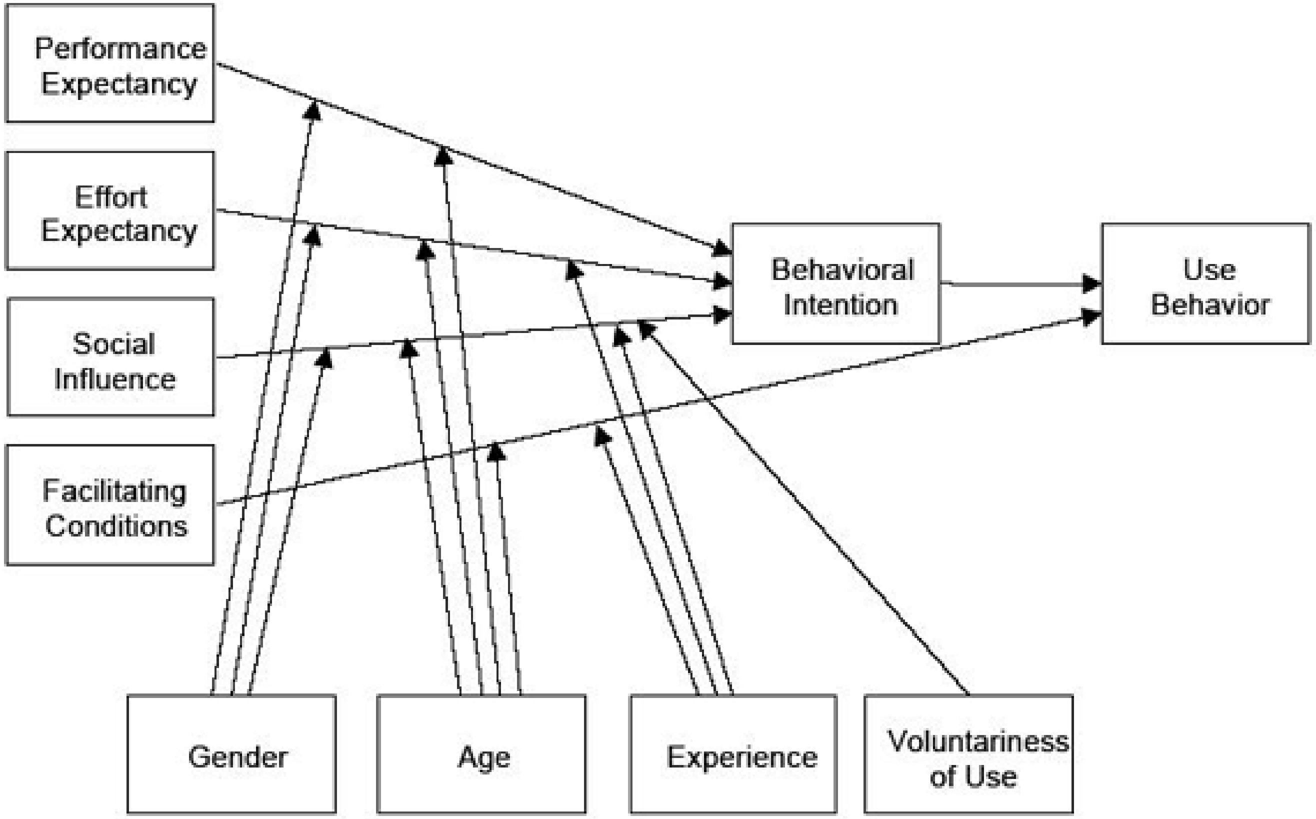
- Biology becomes a digital science
- New technologies with lower costs create both a dangerous situation and an formidable scientific opportunity.
- A system of systems :

  « metadata + collaborative tool + analysis portal »
- Continuum : data centered philosophy

  « Bringing back Biology to the biologist »
- Linked to the research lifecycle
- Acceptance / adoption issues are key issues

UTAUT

Unified Theory of Acceptance and Use of Technology



Venkatesh, V., Morris, M.G., Davis, F.D., and Davis, G.B. “User Acceptance of Information Technology: Toward a Unified View,” *MIS Quarterly*, 27, 2003, 425-478