Data in Forest Research - Challenges and Opportunities

Final Conference 4-6 April 2016, Brussels

Introduction to new faces

Silvia Fluch

fluch@ecoduna.com

(left AIT 2015)

Eva M. Sehr

eva-maria.sehr@ait.ac.at

Stephan Gaubitzer

Overview

- **Data - challenges and opportunities**
- International efforts in data sharing
  - Past: Evoltree
  - Present: Trees4Future
- Conclusions

Data – challenges & opportunities

The most important resource in the 21st century

INFORMATION ⟷ **BIG DATA**

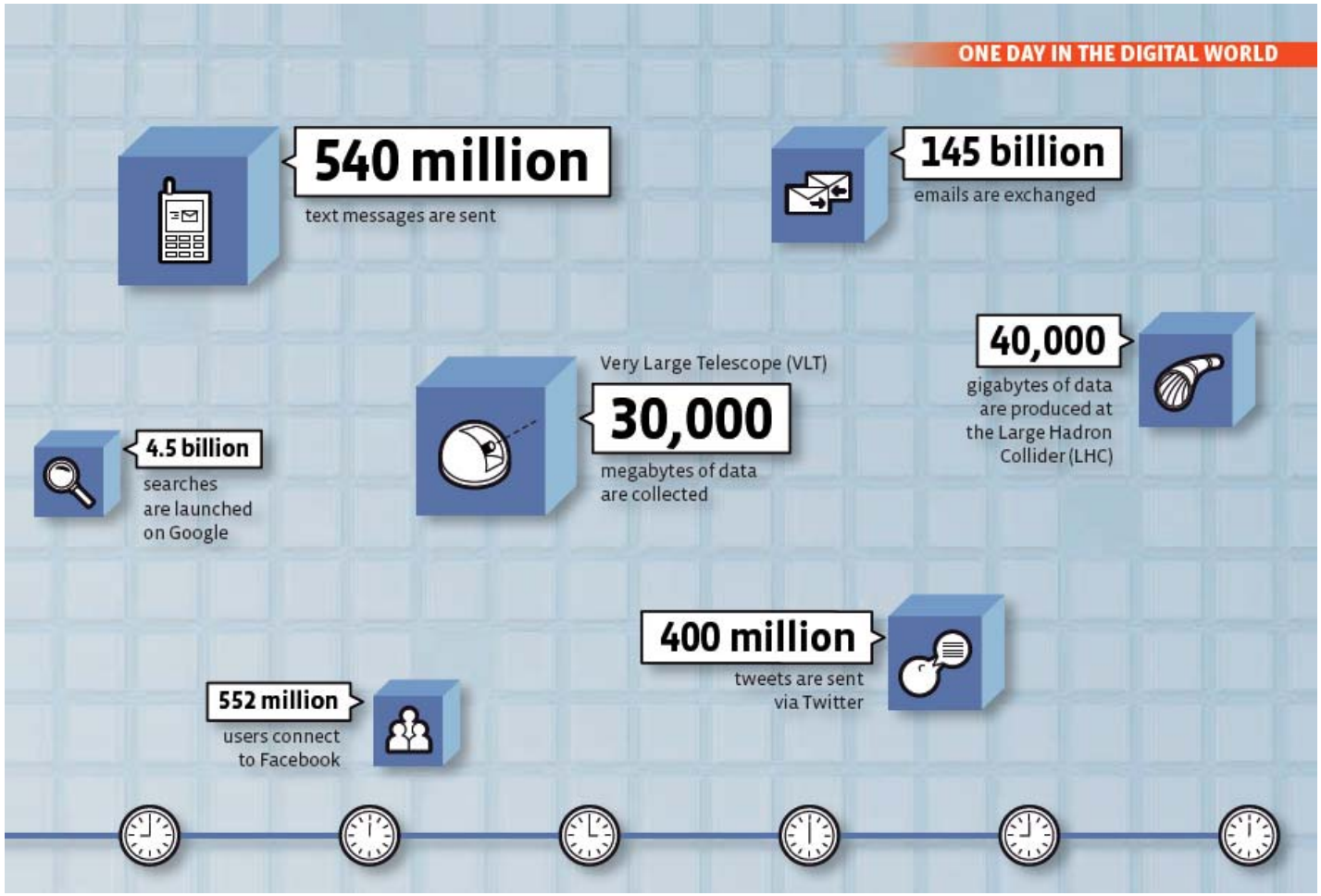
IDC (International Data Cooperation) predicts that

"The digital universe will be 44 times bigger in 2020 than it was in 2009, totaling 35 zettabytes."

1 zettabyte [ZB] = 10^{21} bytes = 1.000.000.000.000.000.000.000 bytes = 1.000 exabytes = 1 million petabytes = 1 billion terabytes = 1 trillion gigabytes = 1 sextillion bytes.

Data – challenges & opportunities

The most important resource in the 21st century



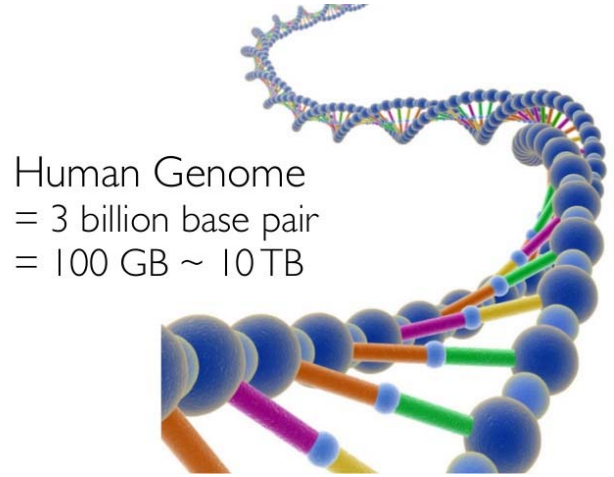
cnrs international magazine, nr. 28, 2013

Data rEvolution in science – 3 V's

- **V – Volume of data**

  Technology
  - Single genes (2000)
  - Single genomes [10 TB] (2008)
  - Landscape genomics (2011)

  Documentation
  - Open source publications
  - Open data (journals)
  - [Open] Specimen records



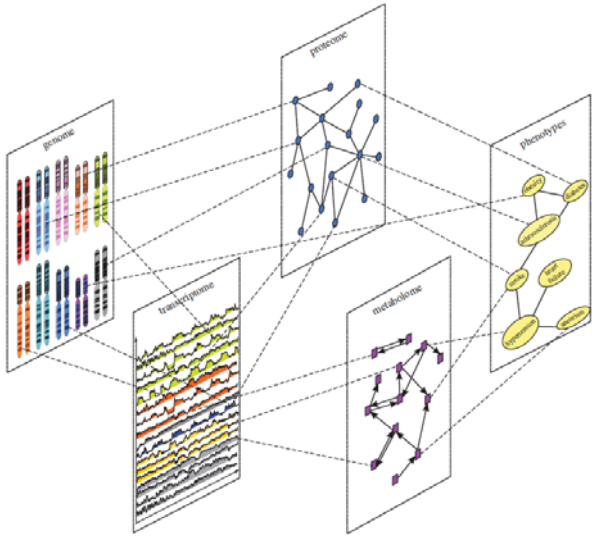
„Human genome sequencing capacity in 2015 is beyond 72.000 TB per year"

(Kovalevskaya et al. 2016, PLOS Biology)

Data rEvolution in science

- **V – Volume of data**
- **V – Velocity of processing data**

  Scientists need high-speed processes to analyze the growing volumes of data.
  - **Supercomputer**
  - **HPC – High Performance Computing**



Gligorijevic & Przulj, 2015

Data – challenges & opportunities

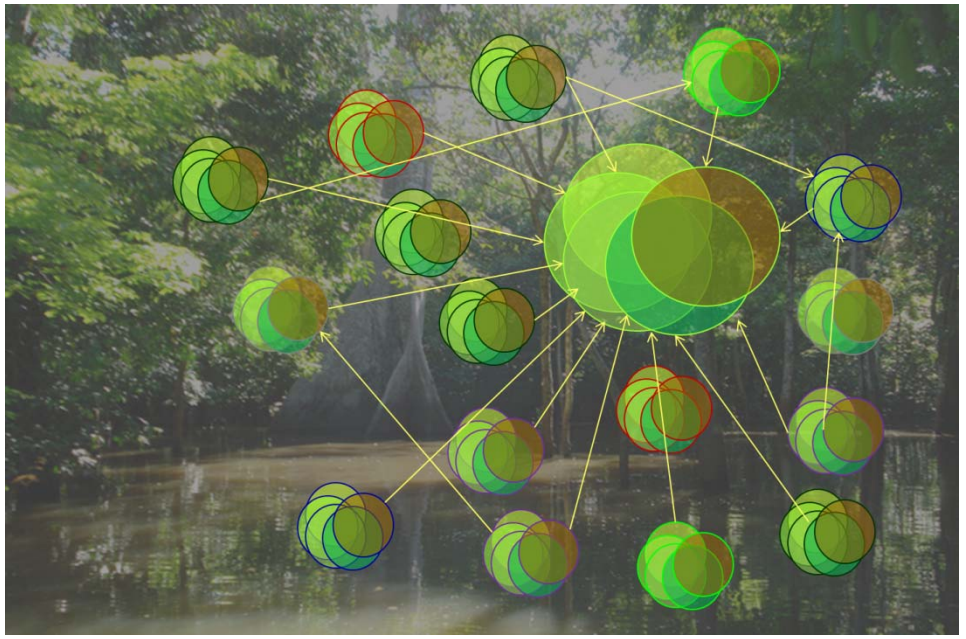
Data rEvolution in science

- **V – Volume of data**
- **V – Velocity of processing data**
  Scientists need high-speed processes to analyze the growing volumes of data.
- **V – Variability of data sources**
  One of the biggest challenges for biologists. Many people from different areas with different data sets are brought together.

Data – challenges & opportunities

Understanding forest trees – data, data, everywhere

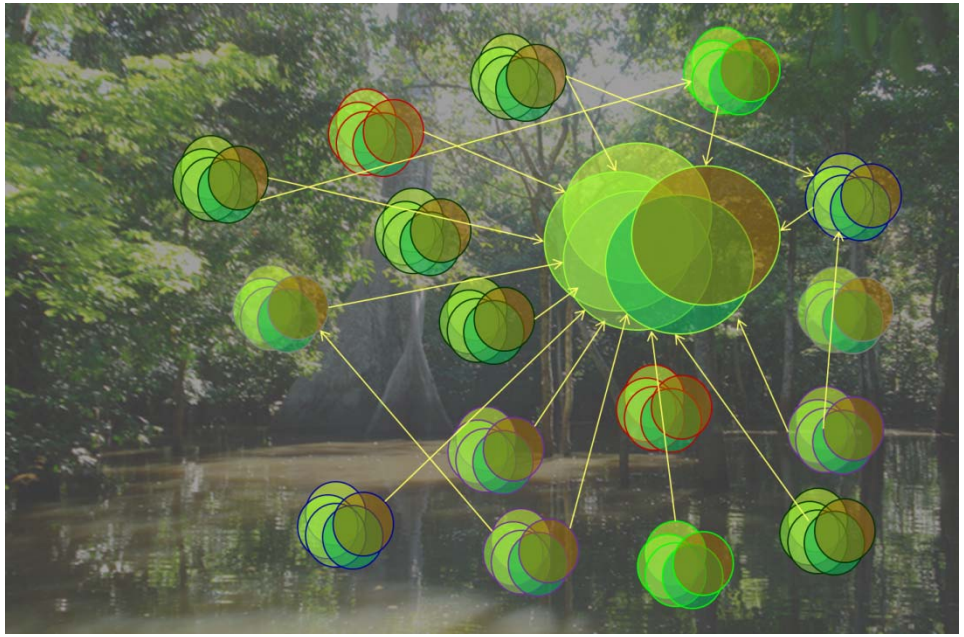
- **Climatic & environmental data** (soil, precipitation, …)
- **Forest tree species (“Omics”-data)**
  - Phenotype
  - Genotype
  - Genome
  - Proteom
  - Metabolom
- **Associated species**
  - Their phenotype
  - Genotype
  - Genome, …



Data – challenges & opportunities

Understanding forest trees – variability of data

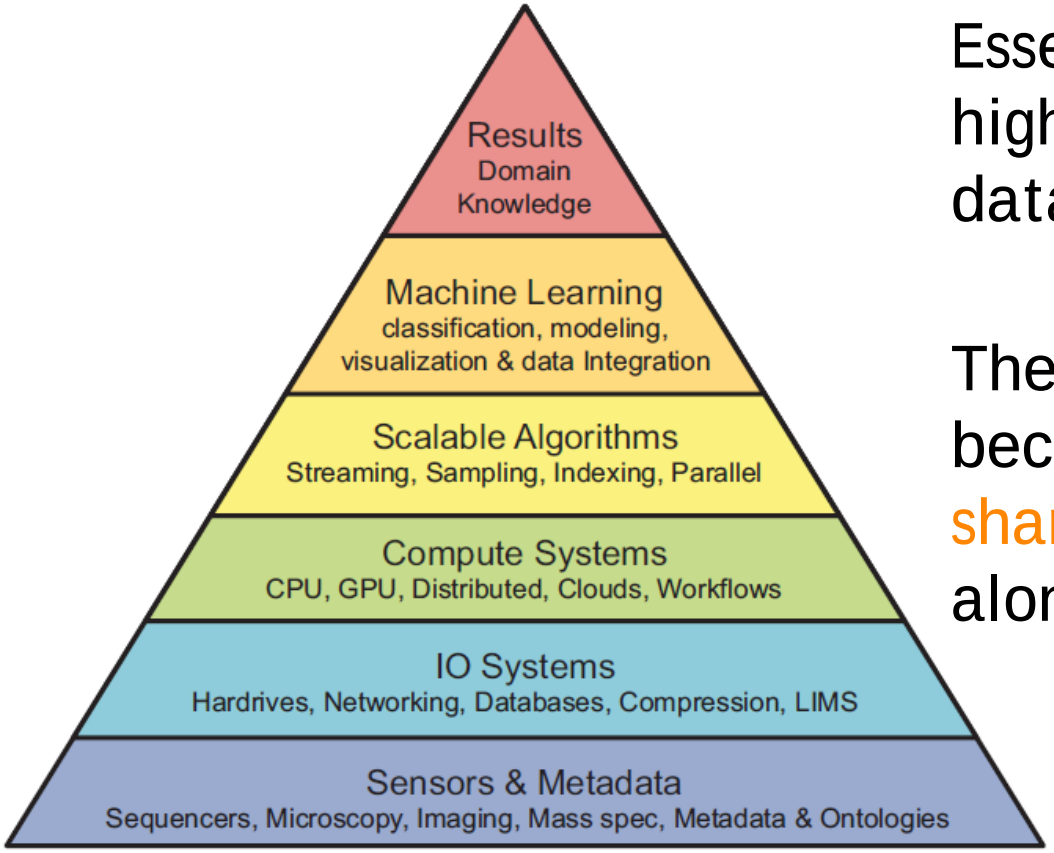
- High quality data sets
- Local (national) level
- Different file formats
- Different languages
- Semantics
- Abbreviations
- …



We are encouraged to facilitate harmonization and synthesis of data across studies.

Data – challenges & opportunities

Understanding forest trees – multilayered approach



Schatz, 2015

Essential prerequisite = high quality of research data through standards!

They guarantee that data become accessible, shareable and comparable along the value chain.

Overview

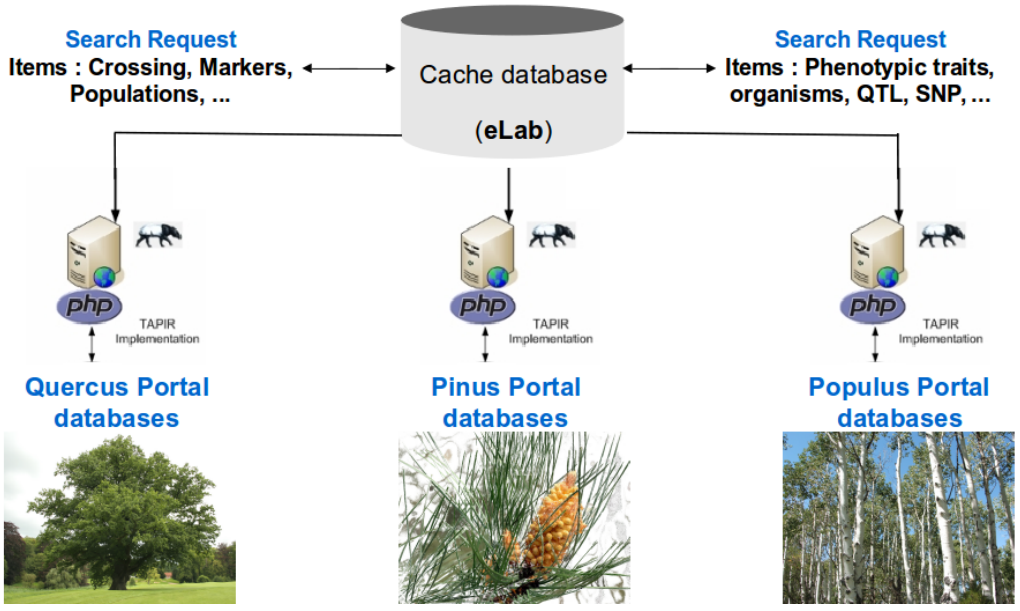
- Data - challenges and opportunities
- **International efforts in data sharing**
  - Past: Evoltree
  - Present: Trees4Future
- Conclusions

International efforts

Past: Evoltree

To maintain and reinforce electronic and physical resources, repositories and infrastructures.

Electronic Lab (eLab)
A centralized search engine, accesses separate databases (passport, phenotypic, genetic and genomic data, …)



www.evoltree.eu

Present: Trees 4 Future

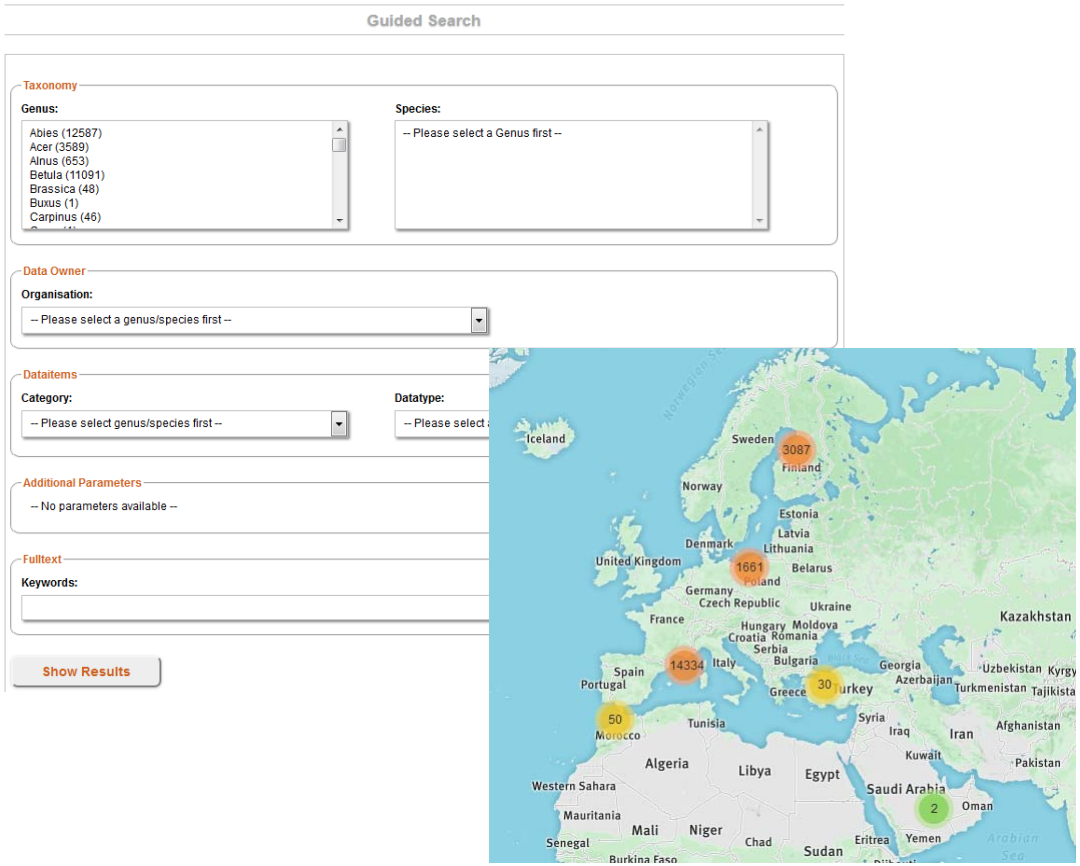
To structure and provide a common access to existing databases from genetics to environmental databases for the benefit of all forest research communities.

20 databases integrated

- AIT GD2 Database
- Candidate-Genes Database
- cMap Pinus
- cMap Populus
- cMap Quercus
- FCBA Database
- GD2 Database
- Gene2Trait Database
- Genfored Database
- GnpIS
- Library & Stackpack Database
- PICME Material Database
- PinusMap Database
- PopulusMap Database
- QuercusMap Database
- SNP Database
- SSR Database
- TREEBREEDEX
- TreePop Database
- Woodtrait Database

Present: Trees 4 Future

Database access



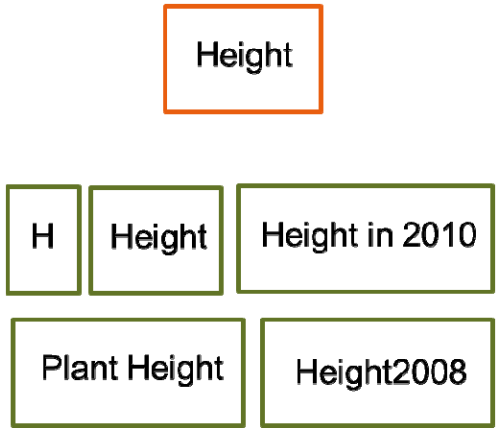
- Overview map
- Full-text search
- Guided search
- Tree view

-> Presentation of databases by Stephan: tomorrow before lunch

Present: Trees 4 Future

Data harmonization, a big challenge

- Data not consistent in all available databases -> currently, every new data set needs to be manually curated
- DB updates need manual interaction
- Same elements might have different names in different databases
- Integration of novel data types
  -> needs manual curation
  -> redesign of existing data bases



Present: Trees 4 Future

Access to standardized information, an opportunity

- Comparability of research results
- Integration of data from various study sites
- Put single studies in a larger geographic/climatic context
- Save money/time for experiments
- Access to different data sets/types for the same sample
- Raw data access for integrated modelling (genetics & climate)
- Access to distributed data sets

Overview

- Data - challenges and opportunities
- International efforts in data sharing
  - Past: Evoltree
  - Present: Trees4Future
- **Conclusions**

Conclusions

- **Resources (data, databases) are available**
- **Integration needs (manual) ‚streamlining‘**
- **Needed: data standardization (guidelines, protocols)**

Ongoing initiatives, e.g. CHARME (Cost Action, Systems Biology data), BD2K (NIH, Big Data to Knowledge initiative, to develop data sharing standards), BioSHaRE (Biobank Standardisation and Harmonisation for Research Excellence in the European Union), OBO (Open Biological and Biomedical Ontologies), etc.

Conclusions

"THAT'S your Ark for the Big Data flood? Noah, you will need a lot more storage space!"