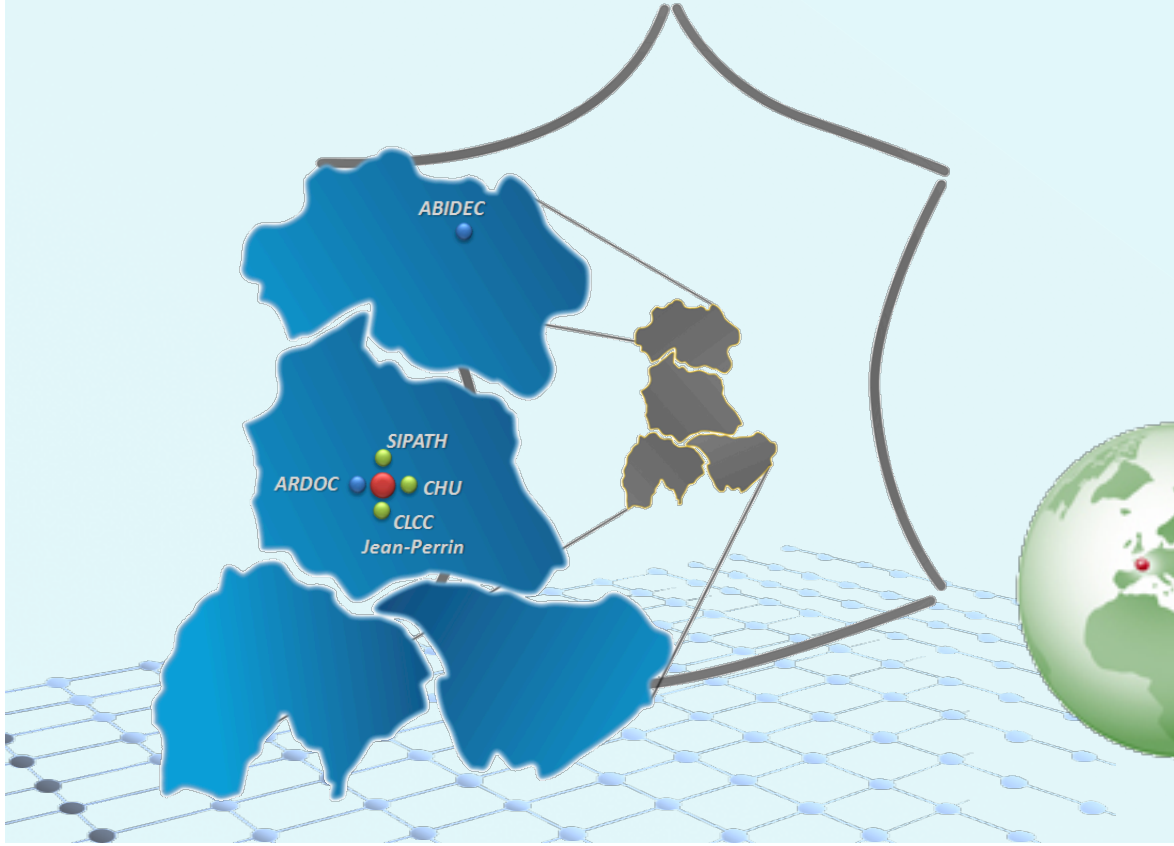
SENTINELLE GRID NETWORK

www.e-sentinelle.org

Sentinel e-health network on grid: developments and challenges

Sentinelle Grid Network

Paul De Vlieger
LPC CNRS/IN2P3
ERIM
vlieger@clermont.in2p3.fr



HEALTHGRID 2010 CONFERENCE PARIS

Breast & Colon Cancers

SENTINELLE GRID NETWORK

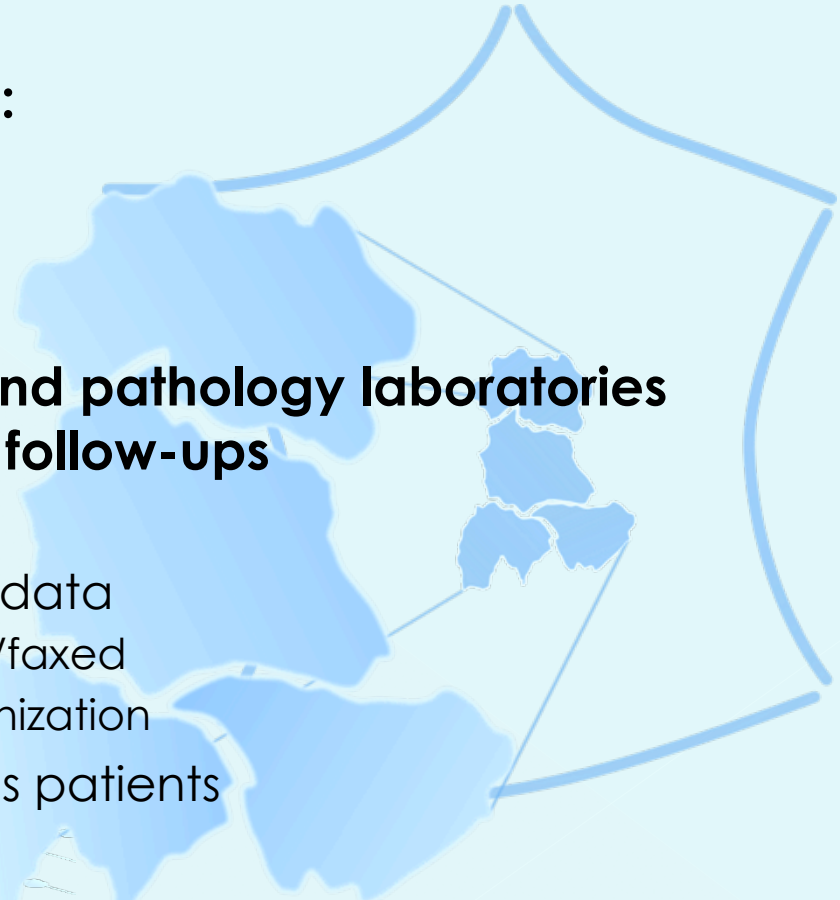
Outline

- **Context: Cancer screening in France**
- **The Sentinelle project**
  - Objectives
  - Architecture
- **Specific issues and challenges**
  - 1 - Data consistency
  - 2 - Patient identification
  - 3 - Data linkage
- **Data linkage results and benchmarking**
- **Conclusion: further steps and future functionalities**

Introduction & context

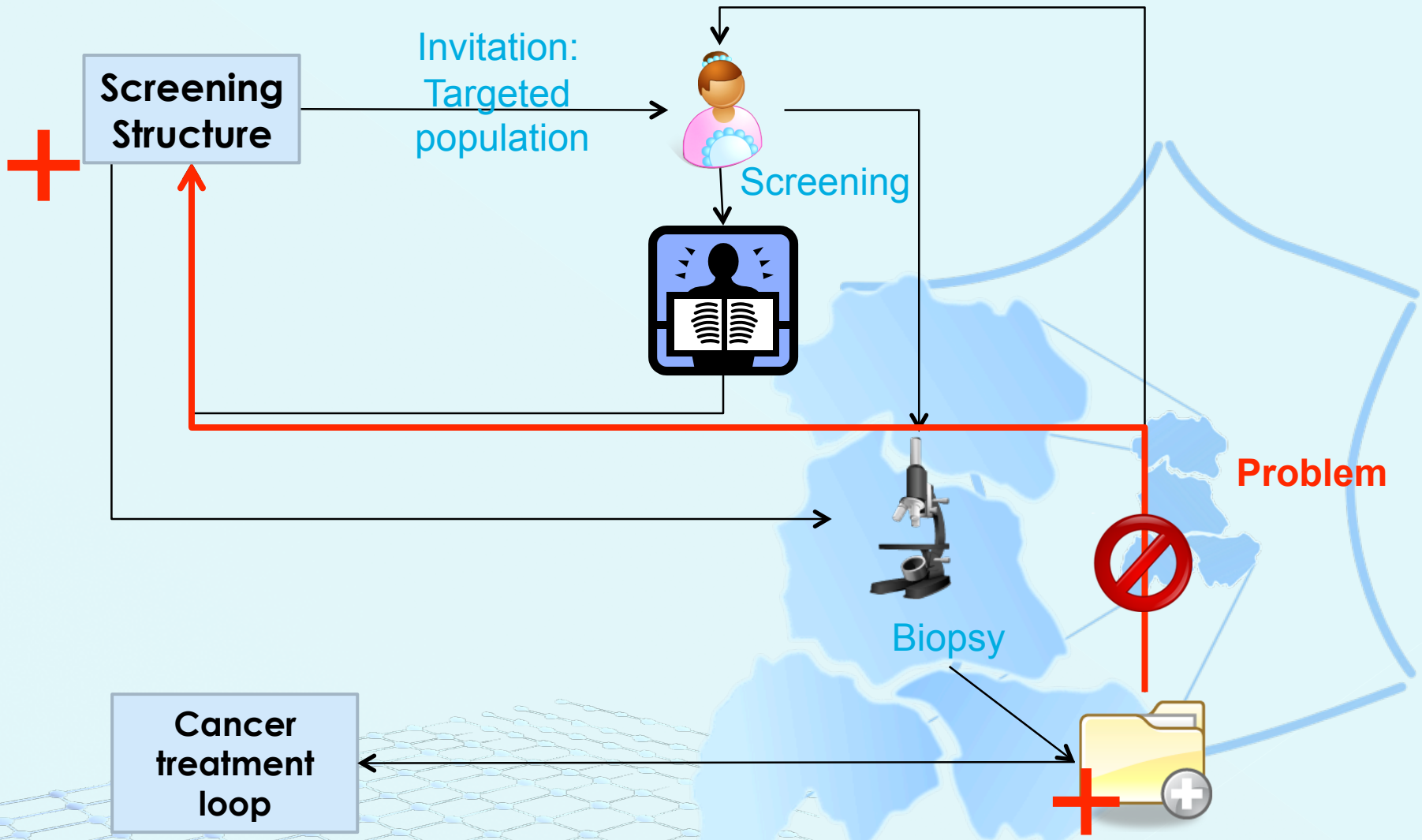
Context

- **Screening**
  - "Earlier is better"
- **3 types of cancers are concerned:**
  - Breast, colon and cervical
  - Easy to screen, easy to cure
- **Cancer screening organizations and pathology laboratories have to exchange data to ensure follow-ups**
  - No electronic exchanges
  - Pathologists refuses to export their data
    - Medical reports are printed/mailed/faxed
    - Re-recorded by the screening organization
  - Data are collected directly besides patients



SENTINELLE GRID NETWORK

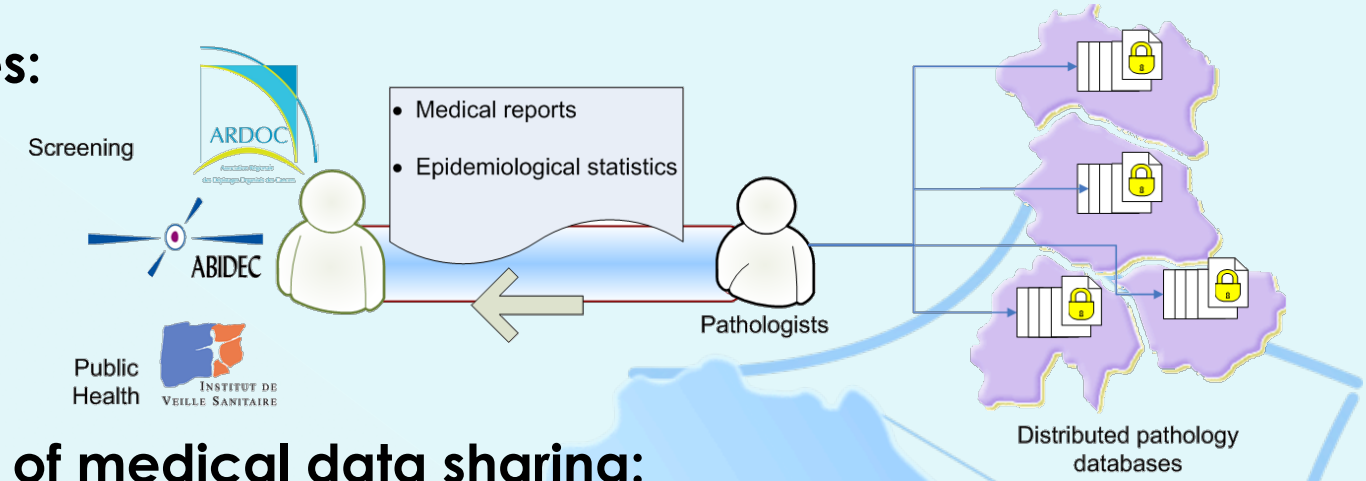
Cancer screening: methodology



The e-sentinelle project

The e-sentinelle project

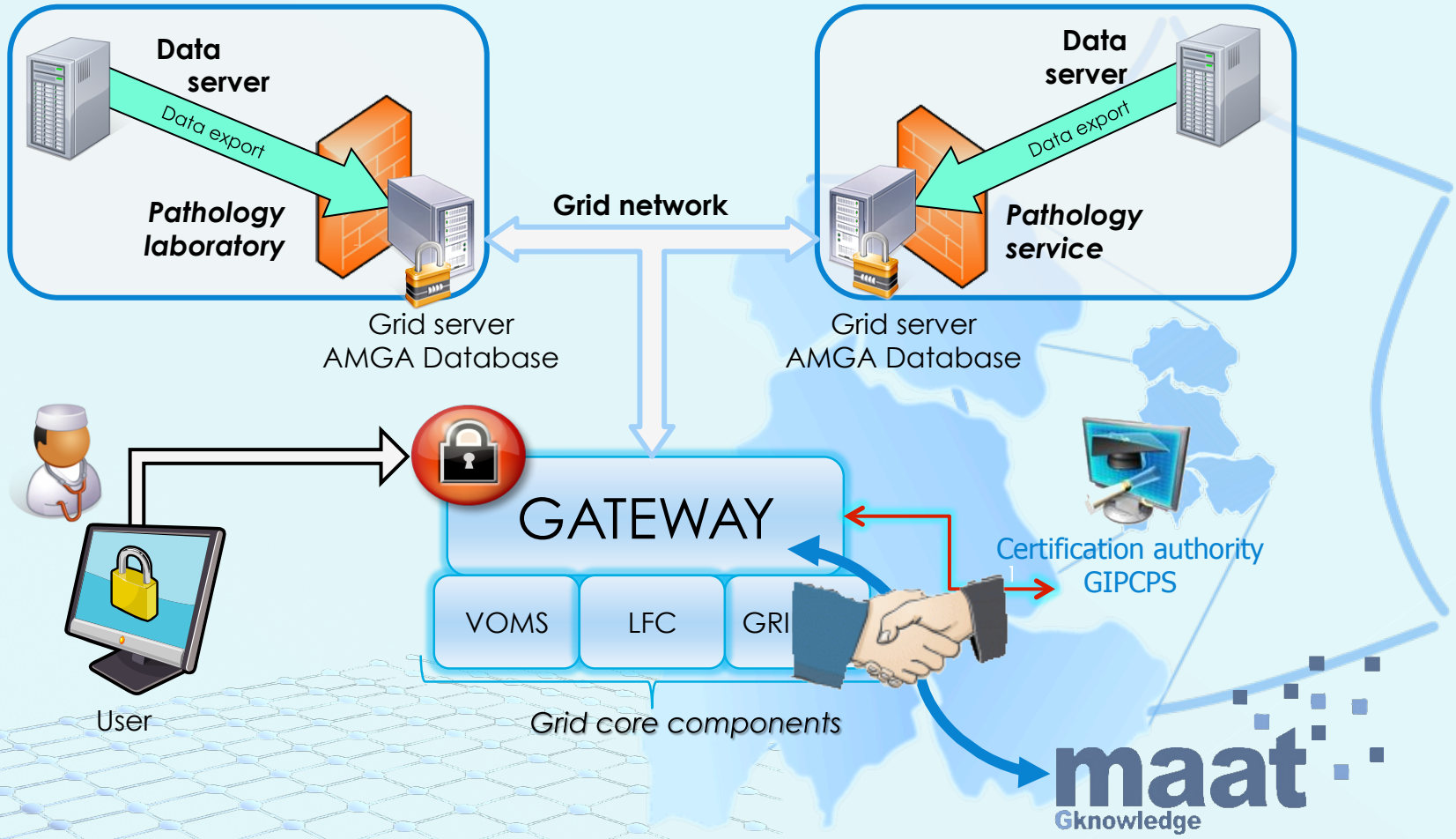
2 main objectives:



- **Improvement of medical data sharing:**
  - Improve and ease collaboration
  - Speed-up treatments
  - Ease follow-up
- **Regional grid network for epidemiology:**
  - Large-scale epidemiology analysis
  - Disease monitoring and alert specification (public health)
- **Health requirements meets grid requirements**

Network architecture

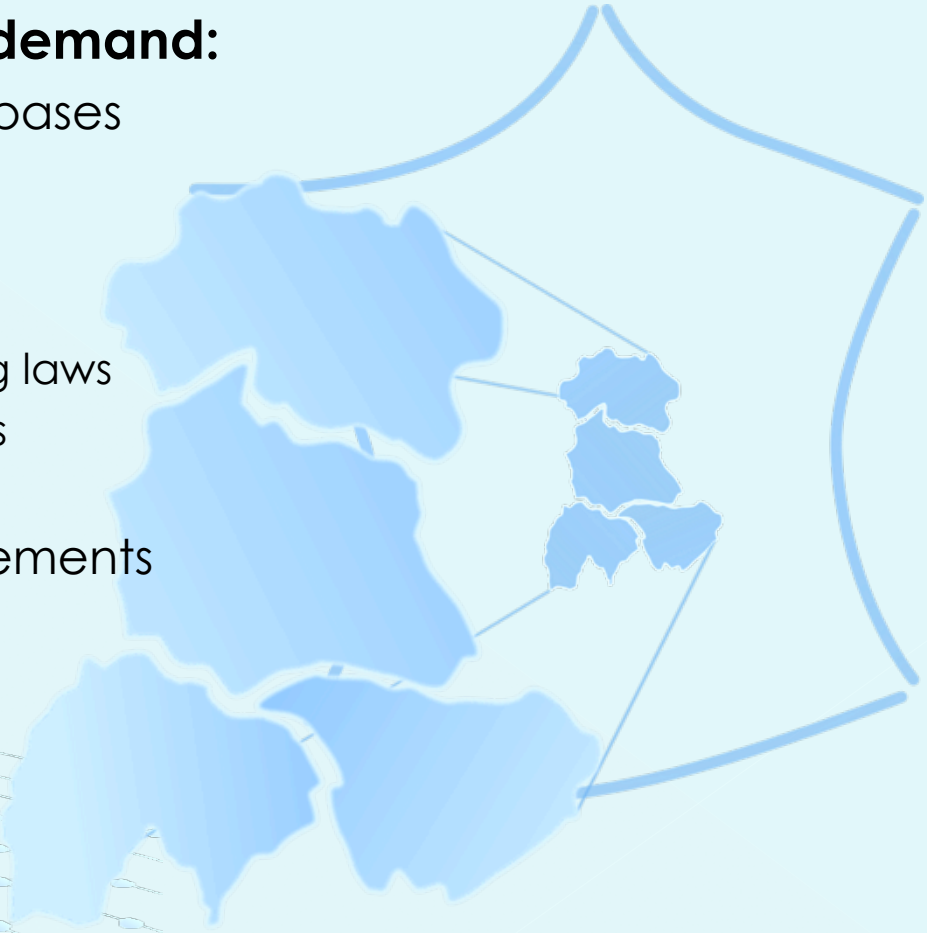
- **A lightweight grid architecture:**



SENTINELLE GRID NETWORK

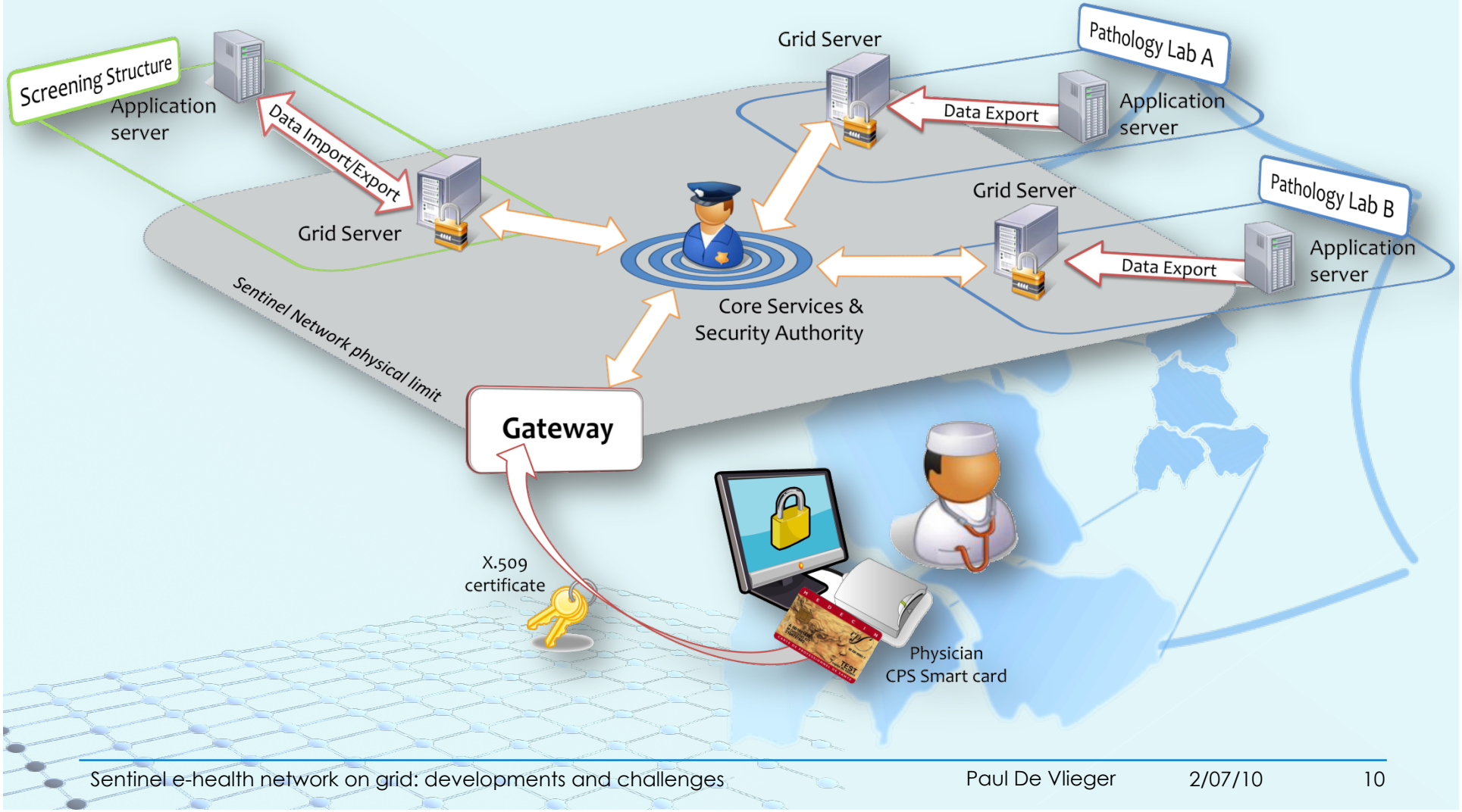
Privacy and security requirements

- **Query medical databases on demand:**
  - No massive extraction of databases
    - Respectful of data ownership
  - Respectful of patient privacy
    - Compliant with data processing laws
    - Use of cryptographic algorithms
  - Guarantying all security requirements
    - Using grid security layers
    - Strong authentication methods



SENTINELLE GRID NETWORK

Project architecture



Specific issues and challenges

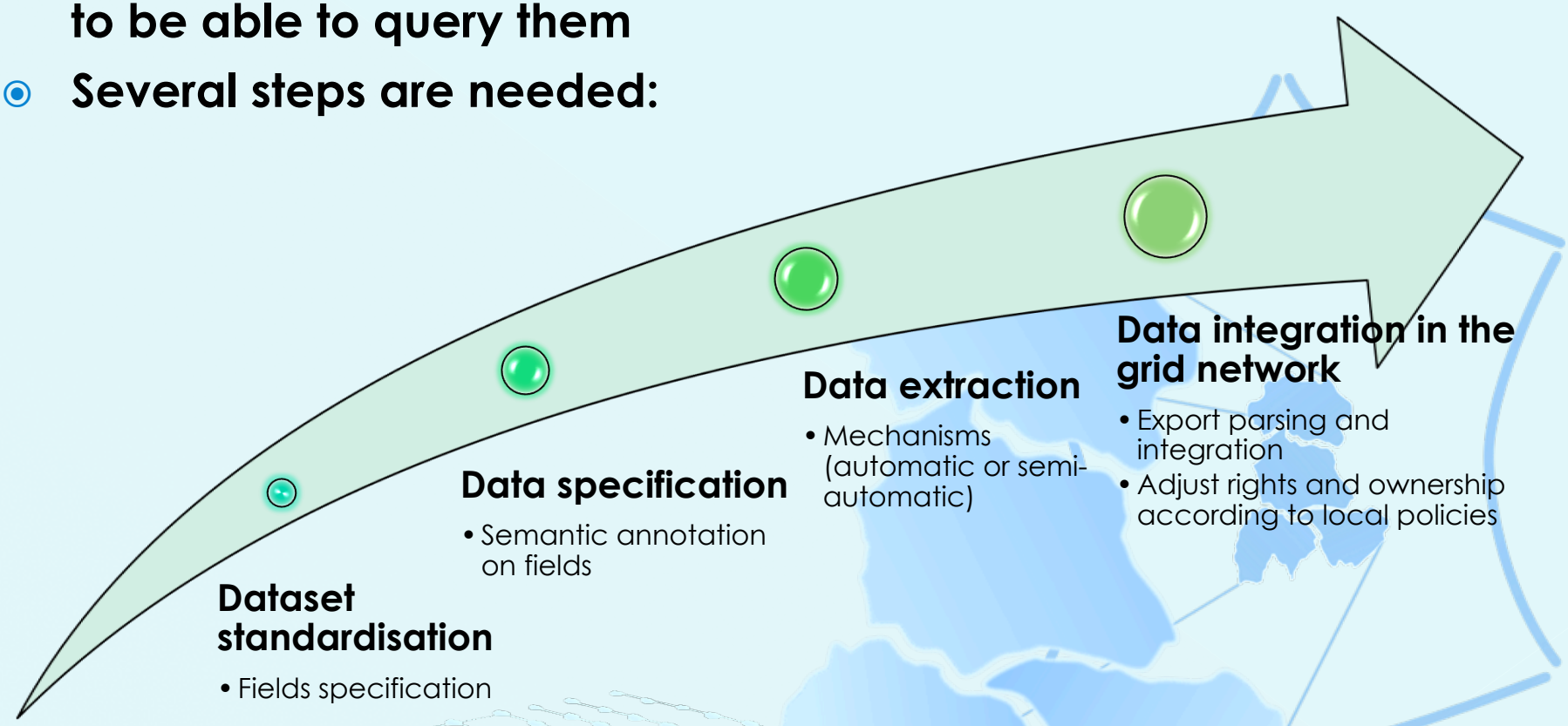
1 - Data consistency

2 - Patient identification

3 - Data linkage

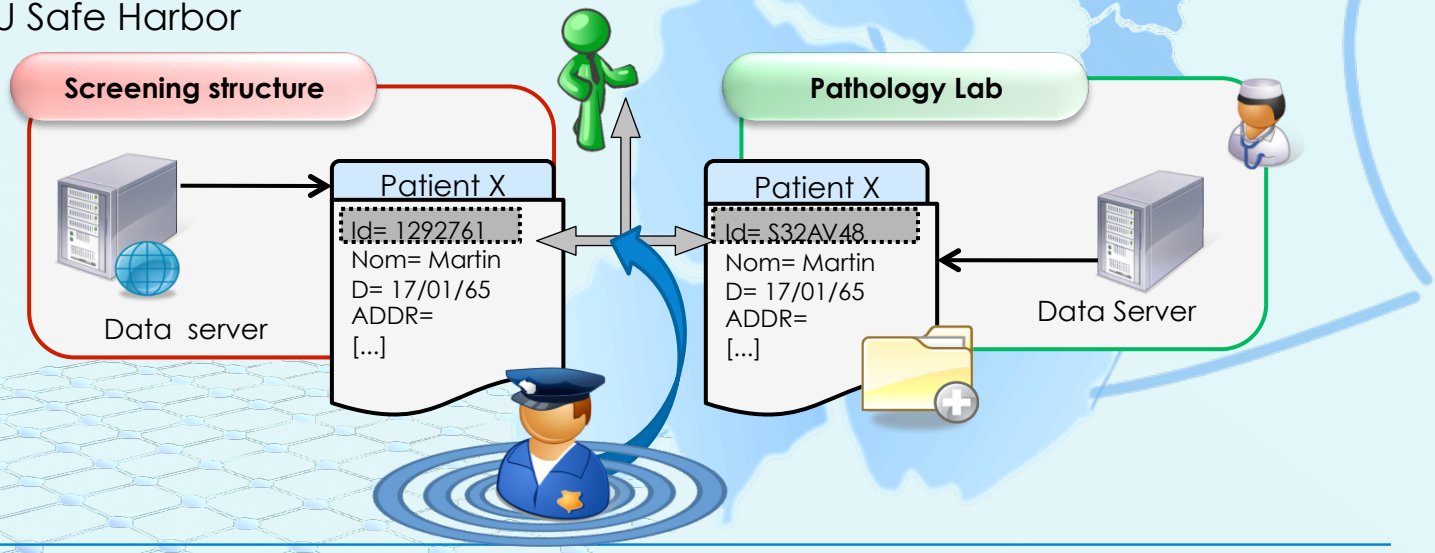
1 - Data consistency

- **Consistency over distributed databases is fundamental in order to be able to query them**
- **Several steps are needed:**



2 - Patient identification

- **Problem:**
  - Identification system is strictly regulated in France:
    - NSS number usage prohibited
    - Not always significant
  - Security constraints:
    - Medical data transfer over a network is regulated by data privacy laws:
      - National laws (French *Informatique & Libertés*, UK *Data protection act*, German *BFDI*…)
      - EU directive 95/46/EC
      - US-EU Safe Harbor



2 - Patient identification

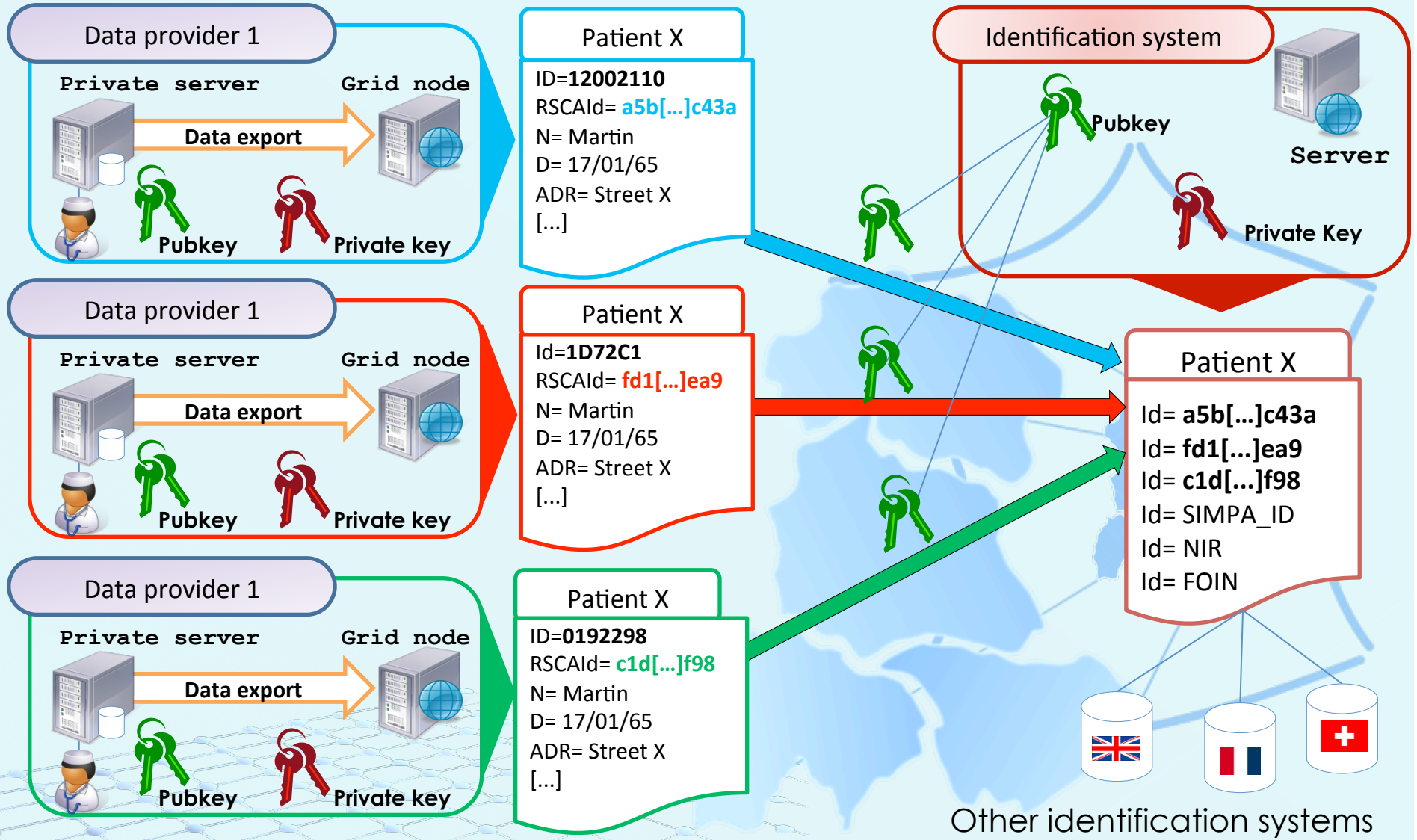
- **Challenge:**
  - Offer a decentralised identification system able to merge distributed patient identities
  - Respect privacy
  - Comply with third party identification sources

- **A new identifier is added to patients for each database:**
  - uuid-based: a 128bit long (as much as possibilities as ipv6 addresses)
  - Randomly generated ex: **`2fffb5aa-07f0-47bf-bb1e-90fda3bada14`**
    - Statistically unique and anonymous

- **Identification system**
  - Consists in merging distributed identities through a centralised server

SENTINELLE GRID NETWORK

2 - Identification system



2 - Patient identification solution

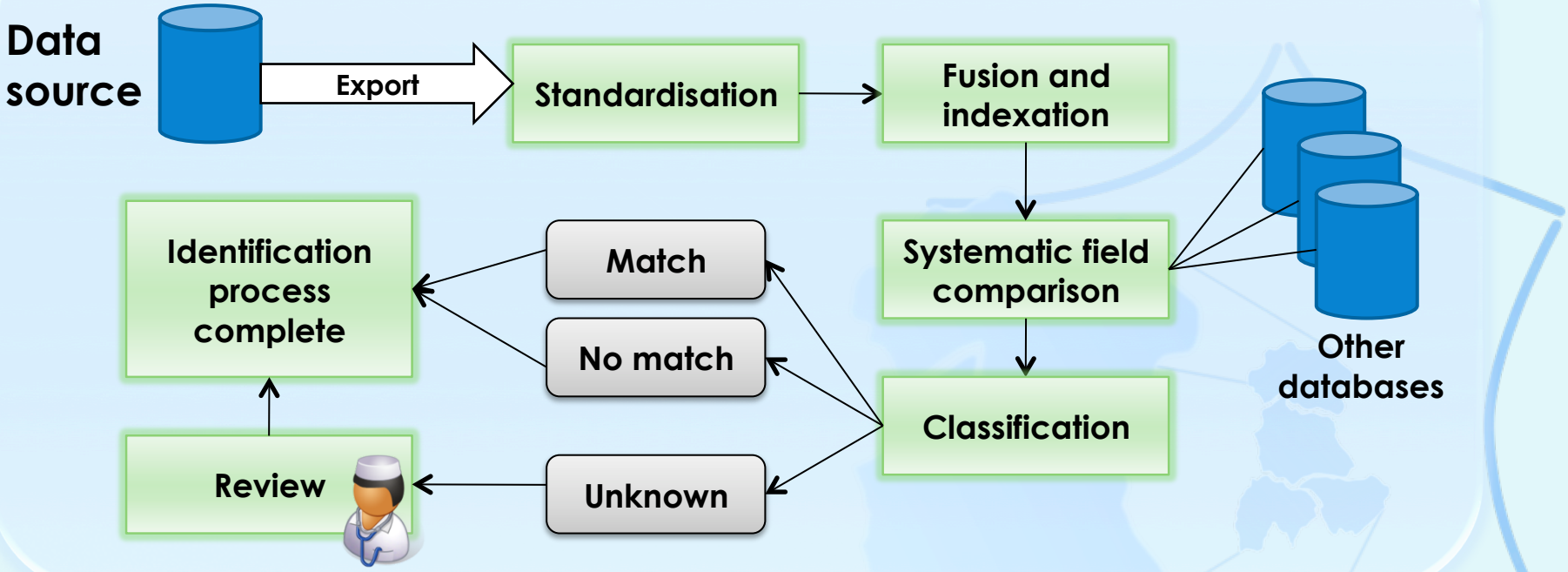
- **Advantages of the solution:**
  - **Scalability**: patient identity can easily evolve
  - **Privacy**: no common identifier exists in distributed databases which remain independent
  - **Identity management**: in case of double identification or two patients with the same number, merging and separating identifiers is straightforward
  - **Interconnectivity**: easy to add another identification source

- **Drawbacks:**
  - SPOF in the central server
  - Lot of encryptions/decryptions
  - 1st phase of identification is costly
  - Need a good data linkage process

3 - Data linkage

- **No common identifier exists => patients have to be compared on their own characteristics :**
  - Name / Surname /Sex
  - Birthdate
  - Address
- **The data linkage process must be as efficient as possible:**
  - Good automatic decision
  - Restricting double-counting (False negatives or type I errors)
  - Avoiding false matching (False positives or type II errors)
- **Data linkage ≠ strcmp in C++**
- **Reasons: Look alike patients**
- **Need high level comparison algorithms:**
  - Measure of similarity between two strings

3 - Identification workflow



- If `score > high_thres` identities are merged
- If `score < low_thres` a new patient is created
- Tricky patient: manual intervention is needed

3 - Data linkage algorithms

2 families:

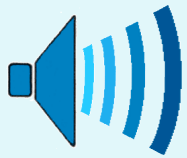
- **String similarity methods:**
  - Jaro-Winkler Algorithm

$$O(\sigma_1, \sigma_2))$$

- **Calculation of a sum of exact characters matches and permutations.**
  - Very efficient for names

- **Rupert / Robert = 0.45**
- **Martha / Matrha = 0.85**

- **Phonetic-based methods**
  - Soundex algorithm
  - Calculates a phonetic codification of strings according to a transformation table:



Class	1	2	3	4	5	6
Letters	B F P V	C G J K Q S X Z	D T	L	M N	R

  - Highly language sensitive
  - Efficient for spelling and phonetic errors

- **Rupert / Robert = 1.0**
- **Martha / Matrha = 0.52**

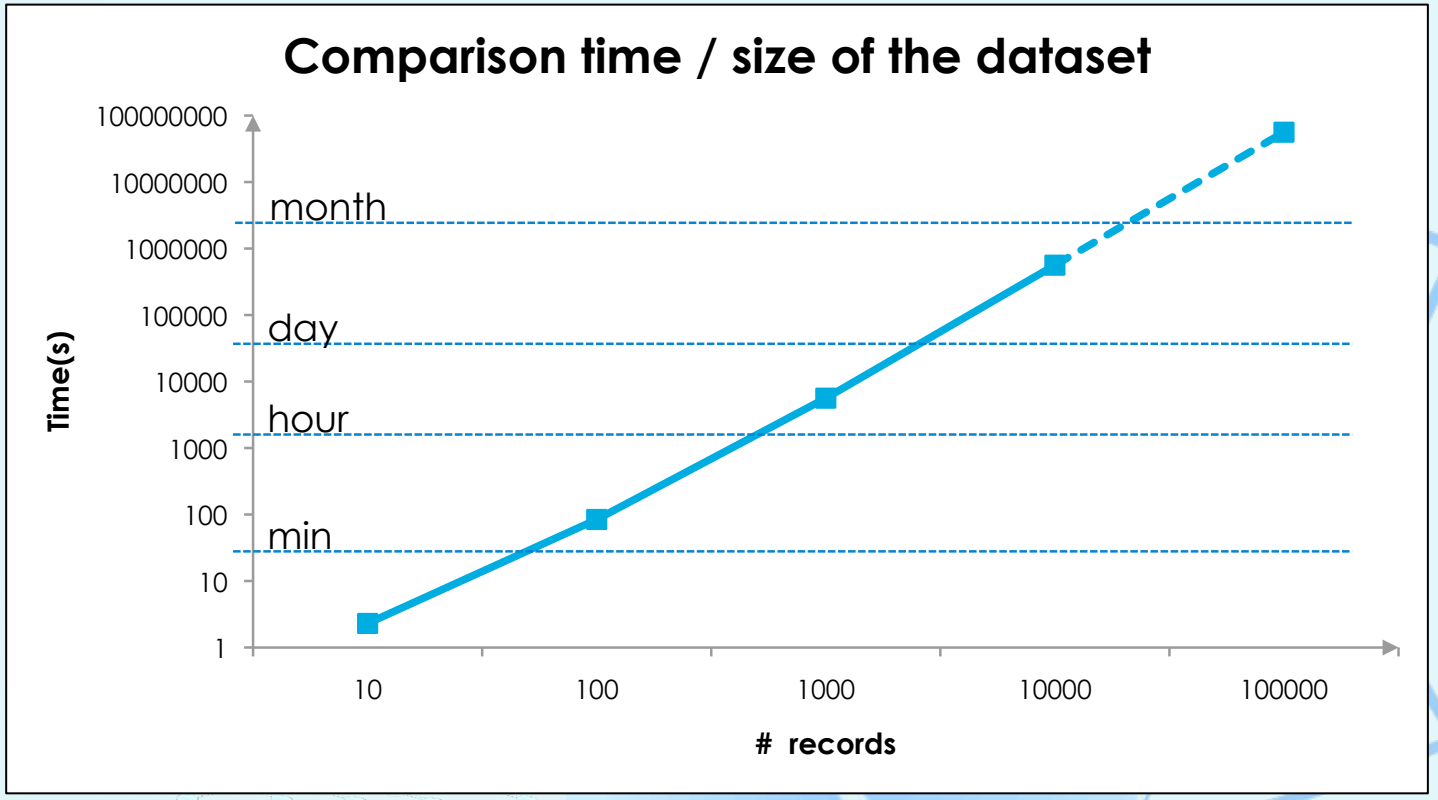
3 - Benchmarking methods

- **Experiment configuration:**
  - Two subsets of 10000 unique records with 12% of joint records
  - Incorporation of random bias in the datasets:
    - Remove/add 0-4 letters
    - Permutations of 0-2 letters with keyboards neighbours (50%) or random letters(50%)
- **Parameters:**
  - Threshold = 0.8
- **Results:**

Field – Method	TP	FN	FP	Result	Accuracy
Last Name – Jaro-Winkler	11.53	1.21	0.06	96.08	90.08
Last Name – Soundex-US	9.33	1.14	0.11	77.75	88.19
First Name – Jaro-Winkler	13.11	2.21	0.09	109.25	85.07
First Name – Soundex-US	10.37	1.93	0.13	86.42	83.43
Address – Jaro-Winkler	9.82	1.72	0.11	81.83	84.29
Address – Soundex-US	7.41	1.72	0.19	61.75	79.51

3 - Efficient methods?

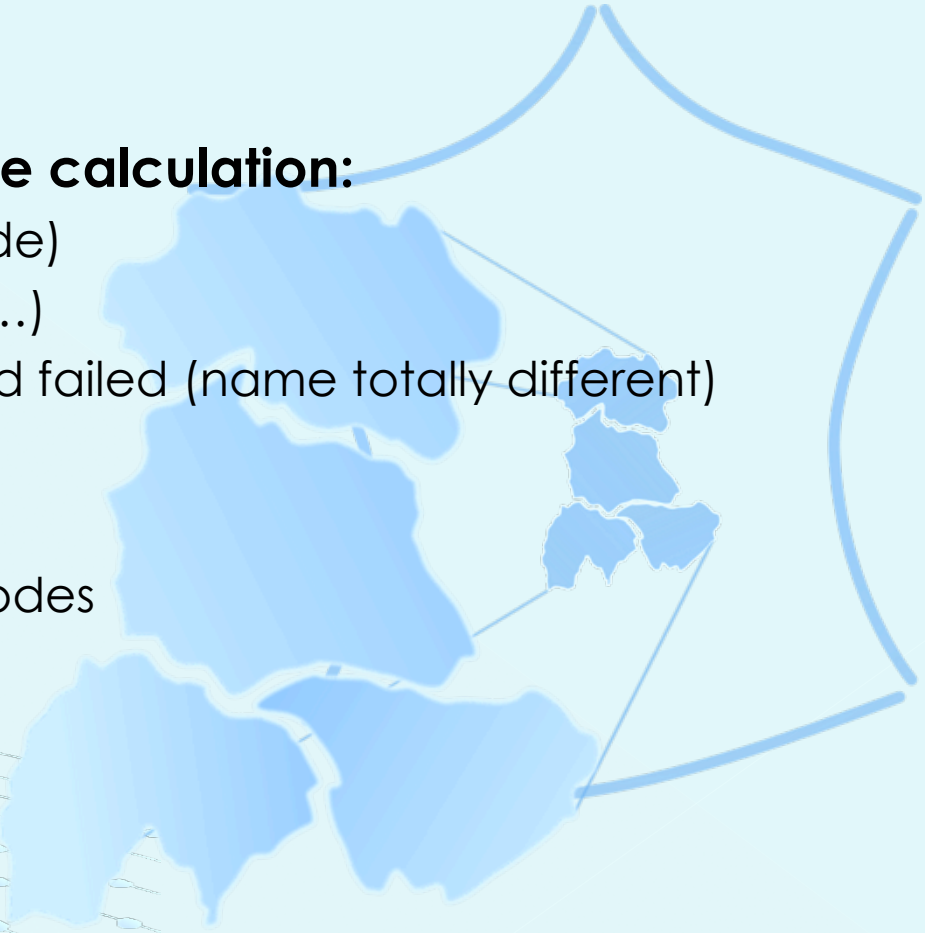
- **Comparison times for 1 dataset and 1 CPU :**



- **Algorithm complexity -> 0(n²)**

3 - Optimization….

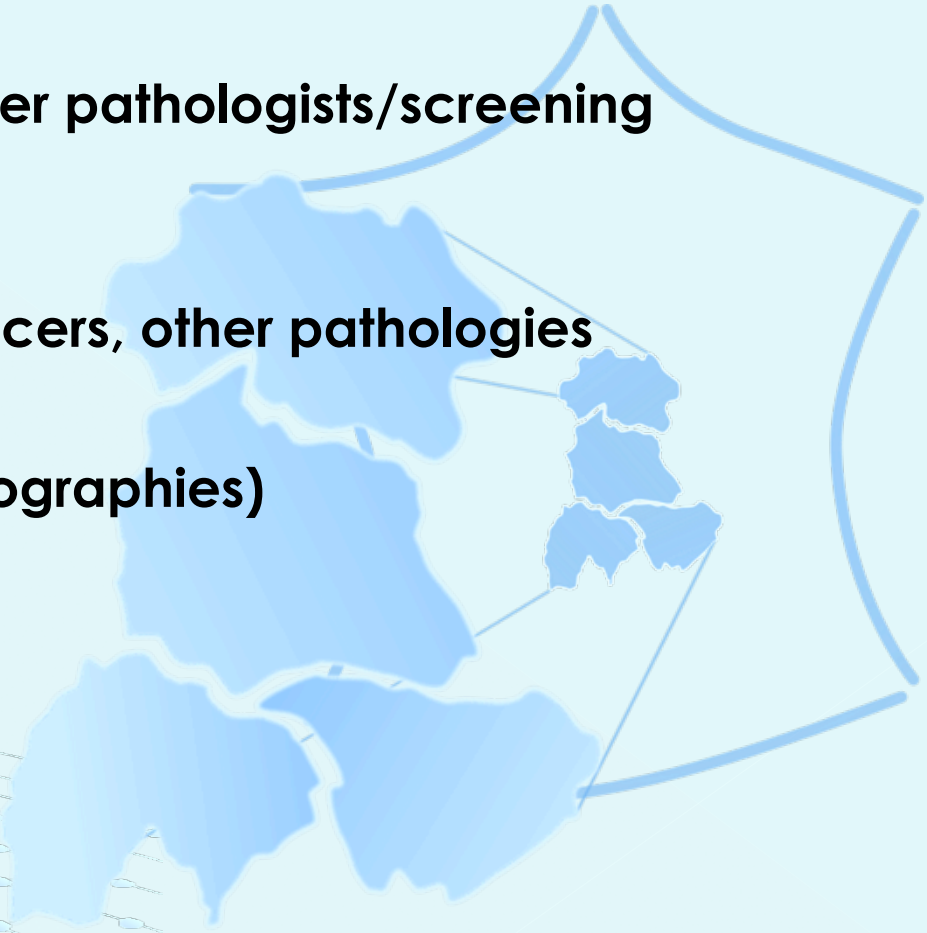
- **Breast Cancer in Auvergne: population of 70000 people**
  - 4.900.000.000 N-to-N comparisons
  - 1month of CPU time
- **Need heuristics to speed-up the calculation:**
  - Geographic restriction (zip-code)
  - Criteria filtering (birthdate, sex…)
  - Stop the process if one key-field failed (name totally different)
- **What else?**
  - Pre-calculation of phonetic codes
  - Use grid computational power



Discussion & Conclusion

Further steps

- **Better integration of the Sentinelle platform inside the medical software**
- **Extension of the network to other pathologists/screening structures**
- **Extension to other types of cancers, other pathologies**
- **Adding medical data (Mammographies)**



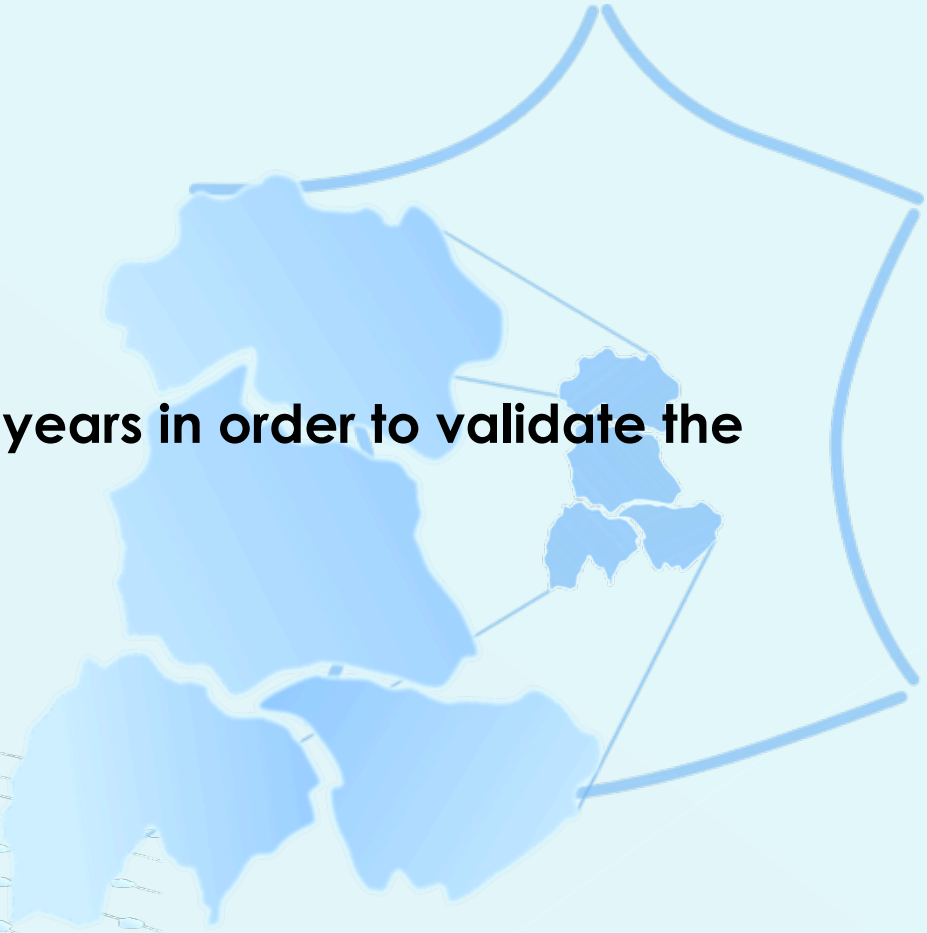
Future functionnalities

- **Complex statistical analysis**
  - Use grid CPU power for cancer cluster investigation (epidemiology)
- **Define and monitor epidemiological alarms**
- **Ease and enhance screening**
  - Multidisciplinary collaborations
  - Real time access
  - Clean duplicates in databases

SENTINELLE GRID NETWORK

Conclusion

- **The e-sentinelle initiative started in 2009 and will enter in production phase soon**
- **Fully grid compliant network**
- **National funding for the next 3 years in order to validate the RSCA**



Thank you

www.e-sentinelle.org