ATF meeting
April 8, 2002

Data Management and security requirements of biomedical applications

Johan Montagnat - WP10

Data

- Public biological data
  - Genomics databases such as SwissProt, TrEMBL, yeast genome…, usually available through web portals
- private biological data
  - Databases owned by private companies, usually disconnected from the outside world. In a grid context they should not be accessible to any system administrator.
- Medical data
  - Every data is protected. Nominative data are only accessible to a very limited number of persons. Images contain both nominative and anonymous data.
- No well established standard format / no standard data structure
  - flat files for biological data
  - DICOM3 for medical images

Users

- Public
  - Anybody an access public databases in read-only mode.
- Owners
  - A limited number of users can modify/update data.
- Patients (individuals)
  - A patient has free access to its medical data.
- Physicians (groups)
  - A physician has complete read access to its patients data. Few persons have read/write access.
- Researchers (groups)
  - Researcher may obtain read access to anonymous medical data for research purposes. Nominative data should be blanked before transmission to these users.

Data management requirements

- Data requirements
  - Very large databases (Tb/database)
  - Long term storage
- Biological data
  - Frequent data updates
  - Data indexing (c.f. BLAST)
  - versioning (CVS-like) mechanism needed to restore older data version
- Medical data
  - Metadata storage (information on patient, acquisition type, medical file…)
  - Access right on a medical department (group) basis
  - Read-only access to raw scanner data, R/W to processed data
  - Record data processing

Use cases

- Web portal access to public data
  - e.g. NPS@ web portal
- Private access to public biological data
  - A private database is compared against a public database. The kind of research should not be known from the outside.
- Private biological data
  - data are only available from, and can only be replicated to, an limited number of resources
- Queries on medical databases
  - A physician retrieves all images from a patient
  - A researcher assess an algorithm on a dataset obtained by querying the data storage system on non nominative metadata

Use cases

- Delegation of credential
  - A physician needs the expert advise of a colleague. He grants access to some image. He could not have grant access to a non accredited user.
- Patient access to its data
  - A patient can query its data, meaning that he can query a database on nominative data, only for data he owns.

Security requirements

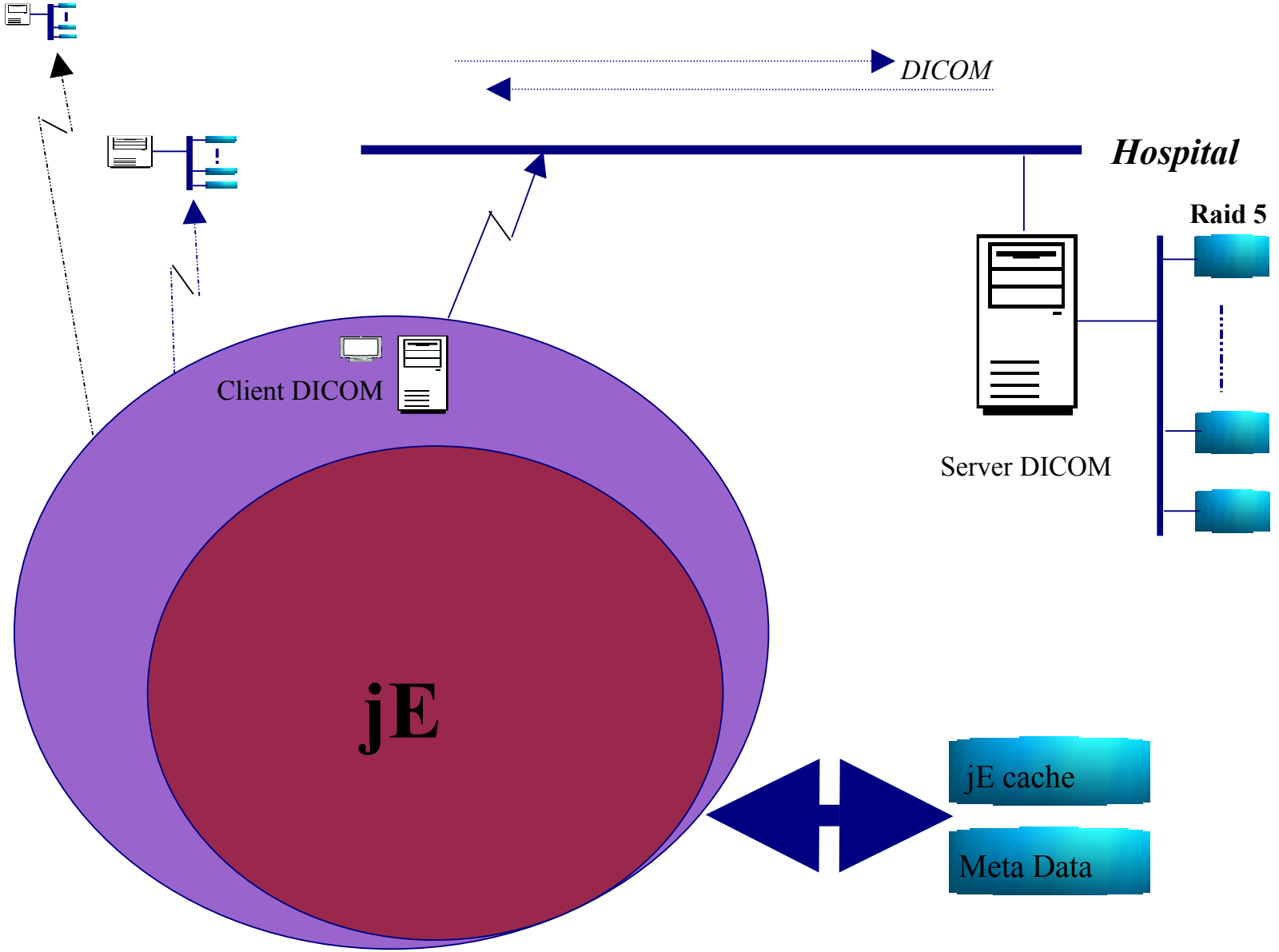
- Security is a central issue for a large scale deployment of WP10 applications
  - sensitive data should not be accessible by third parties (encryption is one mean).
- Data access security
  - Read-only access for most users
  - No read access to private data by any third party user (including system administrators)
  - The replication mechanism should not make replicated data more accessible than masters
  - Quoting D7.5: "it should be difficult for someone to make sensitive data accessible to non accredited users"
  - Sensitive data should not be accessible while being transmitted over the network (secured/encrypted protocols needed)

Security requirements

- Biological data security issues
  - An explicit control of sites where private data may be replicated should be possible
  - A third party should not be able to know what is done with the data (impossible to log other users jobs on the grid)

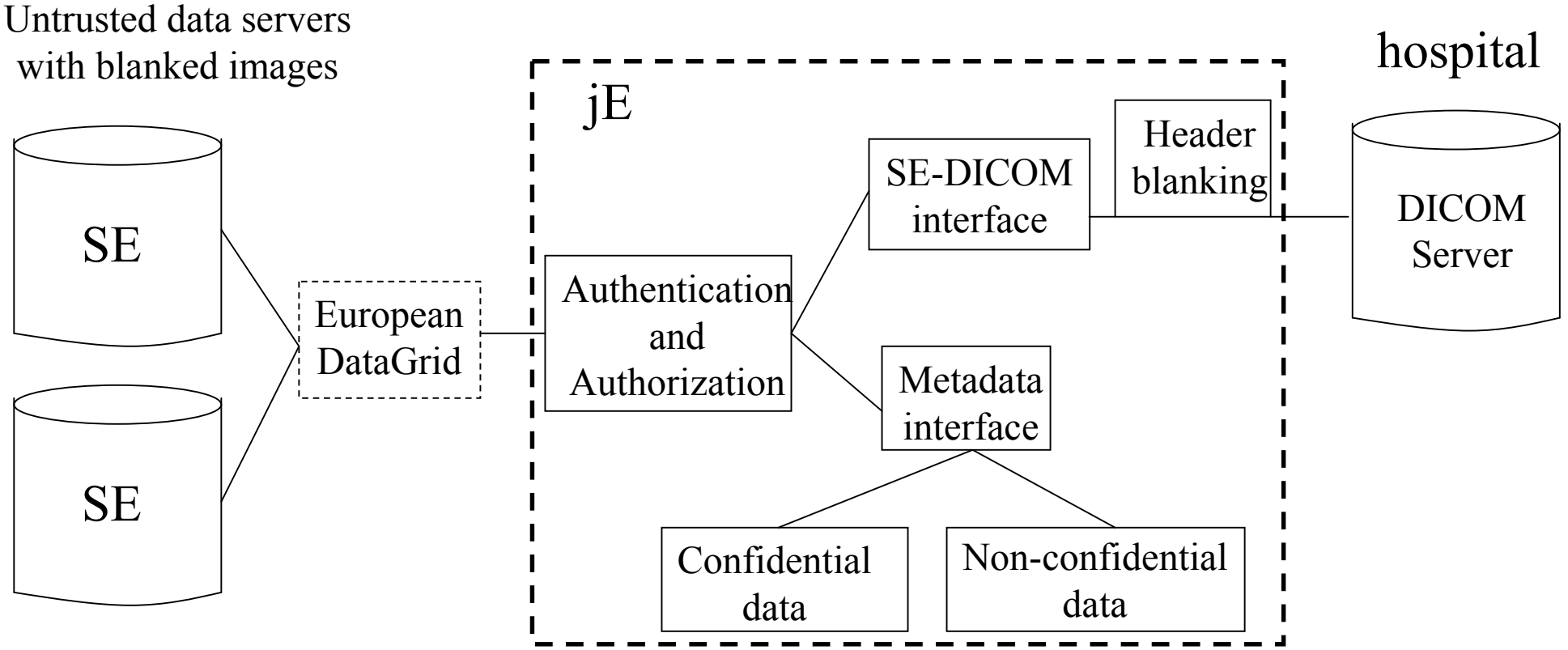
- Medical data security issues
  - Image nominative metadata should be stored in secured/trusted databases
  - Nominative metadata should be encrypted when transmitted over the network
  - Images should not be replicated outside hospital without blanking headers
  - It should be possible for an accredited user to delegate access right for some data to another accredited user.

jE: an interface to medical data servers

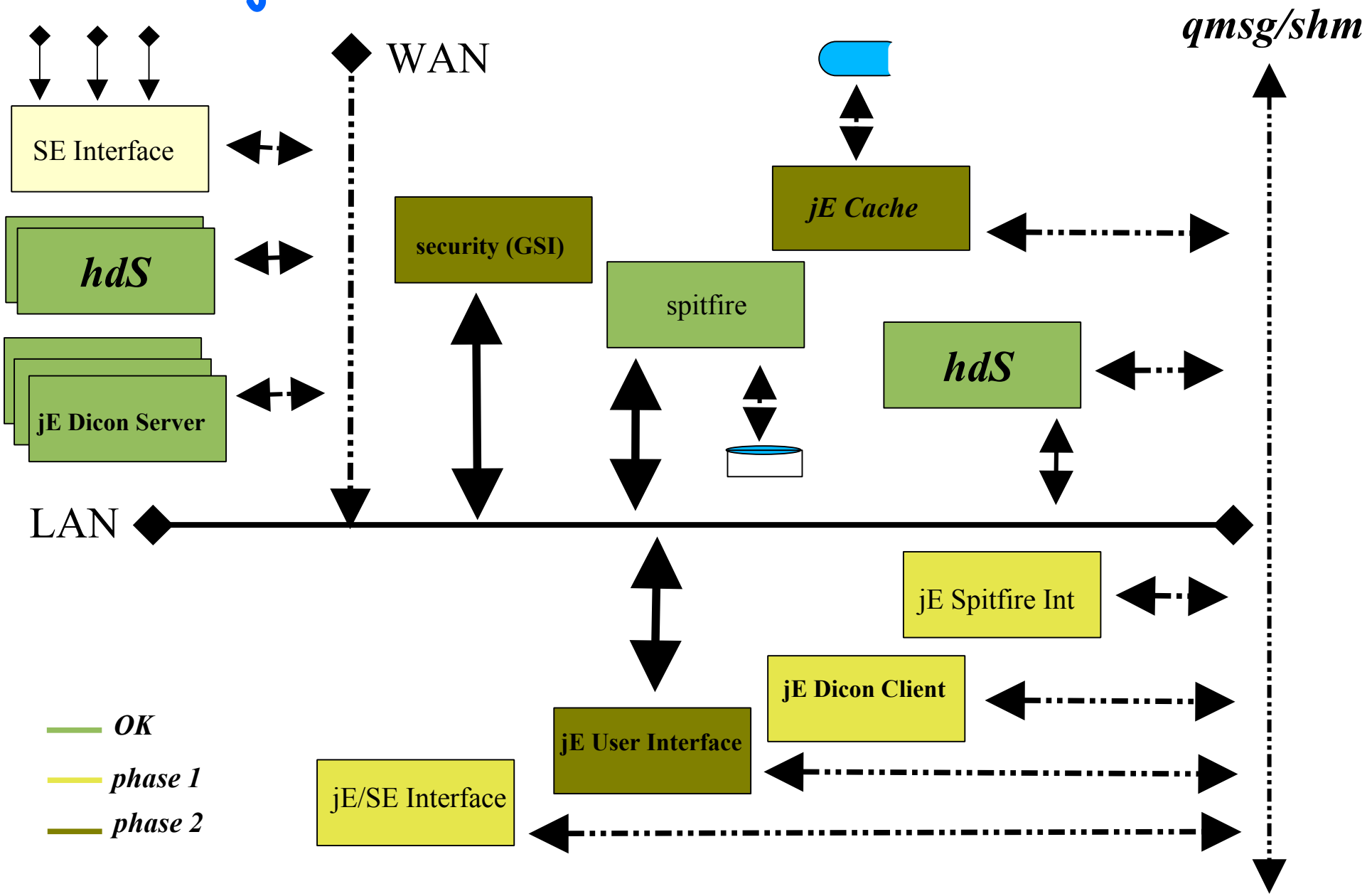


Foreseen medical data infrastructure

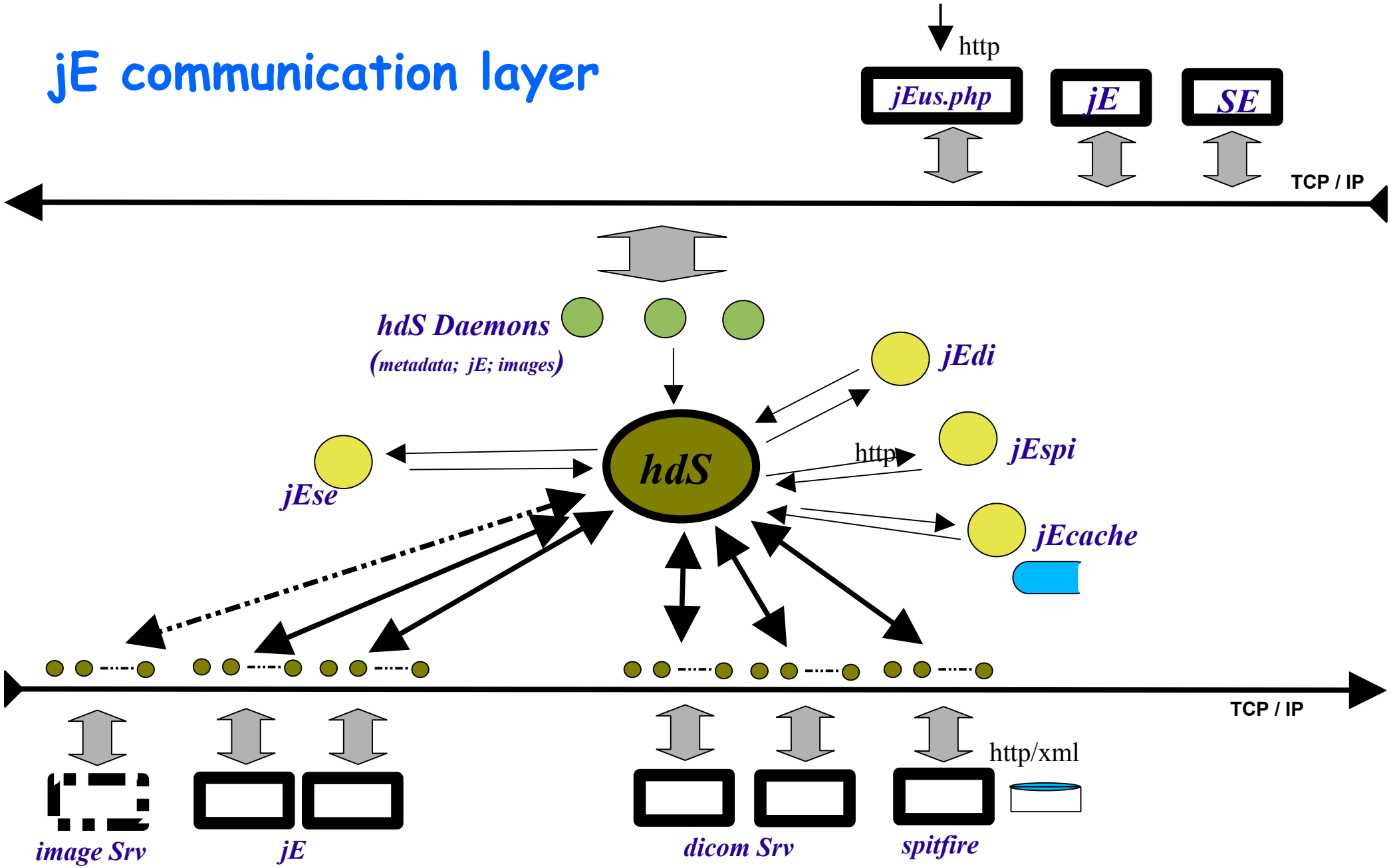
- Split nominative and anonymous data to allow data replication on unsecured sites.



Detailed jE architecture



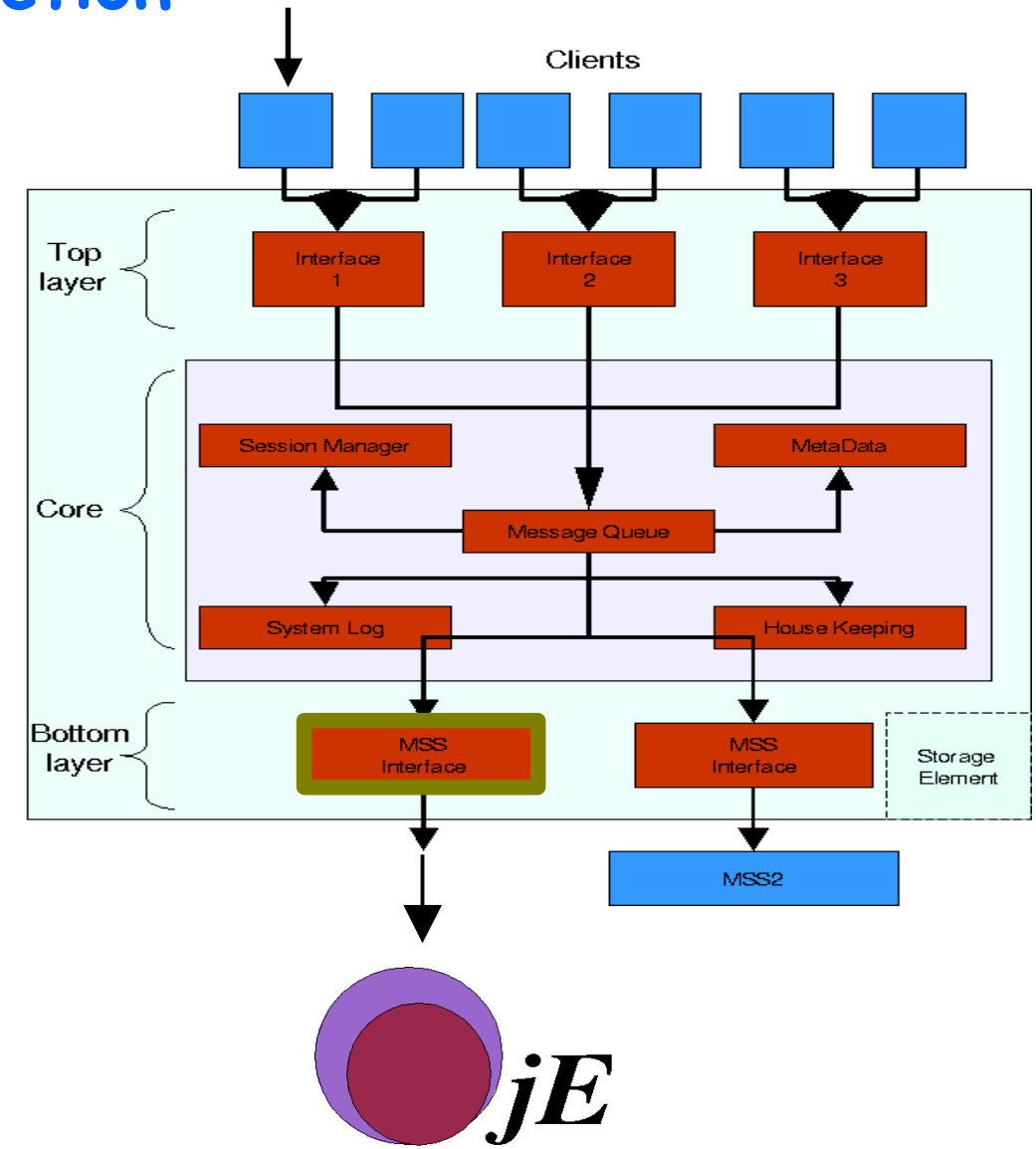
jE communication layer



Questions to WP2

- How to interface the jE with Spitfire ?
  - jwget is a command line tool...
  - Should we wait for an API ?
  - Should we use direct XSQL queries through HTTP ?
  - Which port?
- We need a jE User Interface (to query both local and remote Spitfire databases). What protocol should we use?
  - Http via tomcat server?
    - Could Spitfire tomcat server be reused?
  - Are there any plan to distribute spitfire?
  - Should we use our own protocol?
- Should a jE manage its own RC to make its medical files visible to the outside world?

jE/SE interaction



Questions to WP5

- How to interface a jE as an SE?
  - Should we consider the jE as a MSS?
  - What is the SE interface?
  - What is a SE?
- Which API should use the SE client?
  - RB, RM, JSS, IS, communicate with SEs...
  - Will there be a user application API?
  - Will jE be considered like a user application?