TextGrid und Service Oriented Architecture

GLDV Früjahrstagung 2007
TextGrid-Workshop,
Tübingen, 12.4.2007

www.textgrid.de

Peter Gietz, DAASI International GmbH

Andreas Aschenbrenner, SUB Goettingen

Agenda

- 14:00-14:05 Einleitung
- 14:05-15:00 Marc Küster,TextGrid als digitales Ökosystem: eHumanities meets eScience
- 15:00-15:30 Werner Wegstein, Wie kann TextGrid die Fachwissenschaft verändern – Philologische Edition
- 15:30-16:15 Andrea Zielinski, TextGrid für die Korpuslinguistik
- 16:15-16:30 Kaffeepause
- 16:30-17:15 Peter Gietz, TextGrid und Service Oriented Architecture
- 17:15-18:00 Diskussion über Erwartungen der Fachwissenschaften an TextGrid

Overview

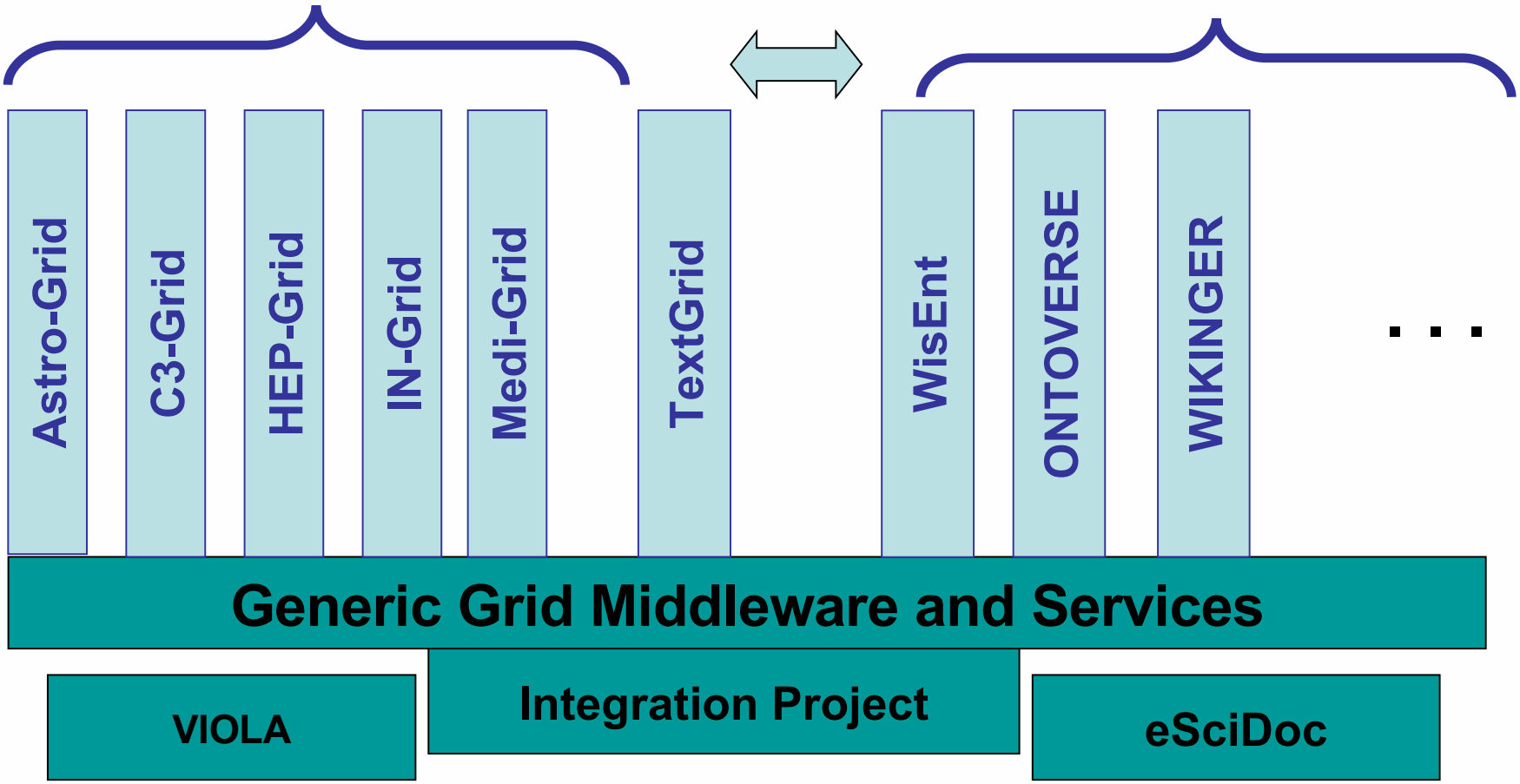
- **TextGrid, Humanities and Grid Computing**
- Objectives and requirements
- Technology

BMBF e-Science Programme

- 2005 - 2009
- 100 institutions
- 100 Mio Euros
- focus areas:
  - e-Learning (June 2004)
  - D-Grid (August 2004)
  - Knowledge Networking (November 2004)

Federal Ministry
of Education
and Research

D-Grid + Knowledge Networking



What about sustainability?

- Different models for a post government-funded grid infrastructure for eScience and eHumanities
  - single researcher has to pay for each job
  - DFN model: universities pay a flat rate
- The second seems to be more realistic

Humanities Computing

- From the 1960s onwards humanities use Computer for their research
- Tools can be summarized into 3 Groups:
  - highly specialized Project specific Tools
  - general purpose text processing tools
  - modern scripting languages (very general purpose)
- An important step has been the standardization of the scholarly text markup language TEI

Web Services and Grid Computing

- Web Services Technologies have arrived in Humanities Computing, e.g.:
  - Tapor (www.tapor.ca)
  - DAM-LR (www.mpi.nl/DAM-LR)
- Science has established Grid Computing:
  - transparent access to Computing and Storage facilities in the network
  - used in, e.g. HEP to cope with Petabyte of data and complex data analysis algorithms
  - WS based standards are already there (OGF)

eScience and eHumanities

- eScience
  - a new form of network and Grid-based scholarly work and collaboration
  - allows for innovative research designs
- eHumanities
  - equivalent of eScience in the field of humanities
  - deployment of Grid-based infrastructure and network collaboration within Virtual Organizations
  - “e” also stands for enhancing, extending and enabling

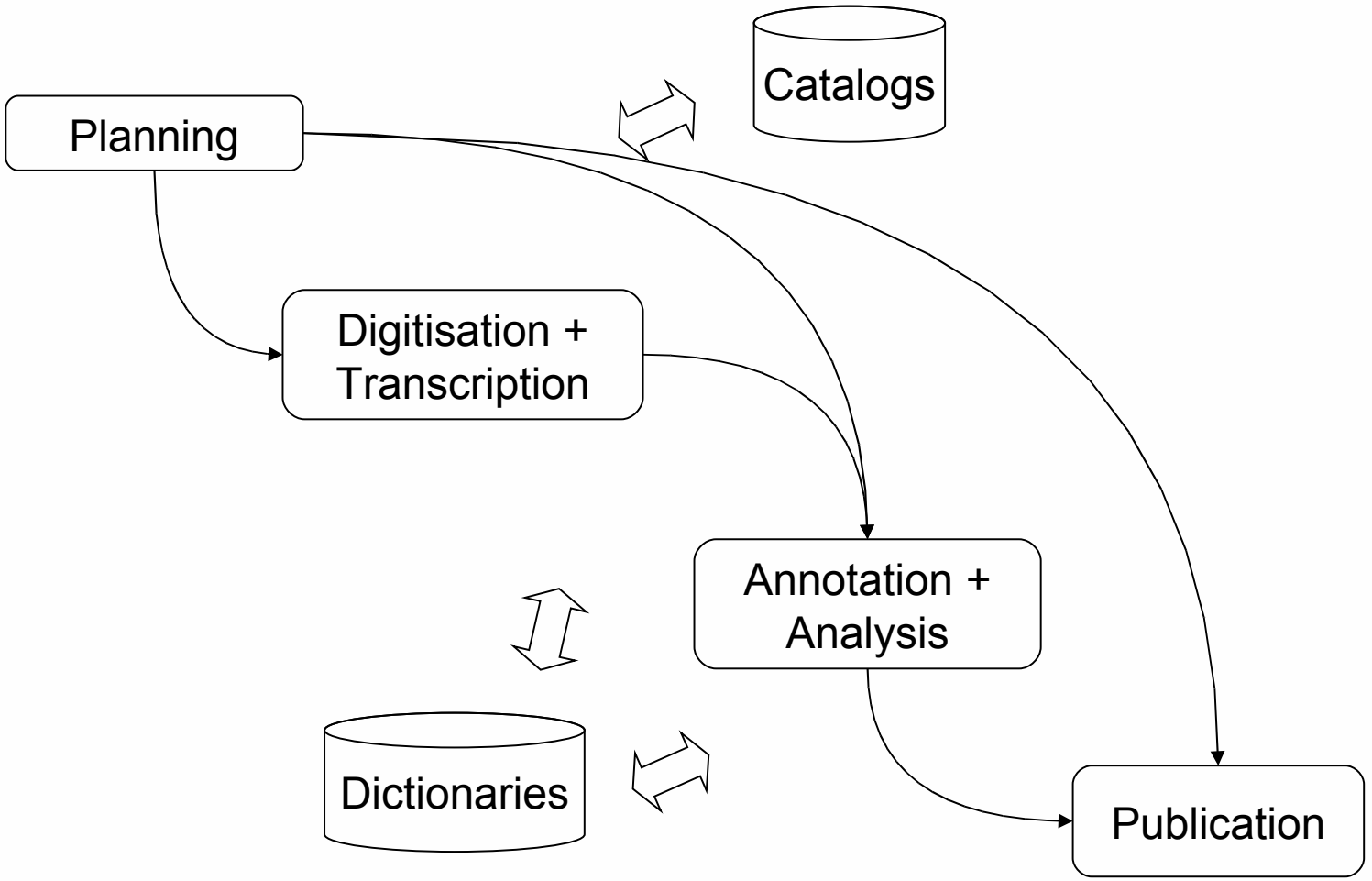
TextGrid and eHumanities

- TextGrid establishes eHumanities in the field of scholarly text processing
  - deploying Grid and humanities Computing standards
  - providing distributed digital texts and processing tools for answering traditional and new questions with empirical methods
  - promoting research in a networked and interdisciplinary environment
  - developping a modular and open platform for scholarly text processing

Overview

- TextGrid, Humanities and Grid Computing
- **Objectives and requirements**
- Technology

workflow


Catalogs
Planning
Digitisation +
Transcription
Annotation +
Analysis
Dictionaries
Publication

TextGrid goal

an open, adaptable, extensible infrastructure

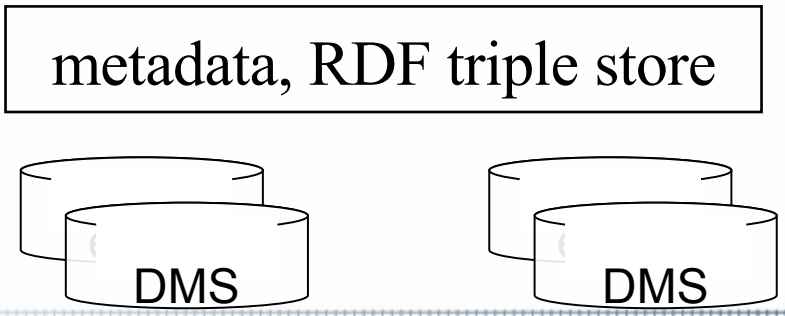
- combine isolated text corpora for re-use and collaboration
- enrich content with supplementary resources (dictionaries, catalogs)
- modular, to re-combine services for tailored tools
- foster automation on a semantic level

metadata

- bibliographic (title, author, ...)
- administrative (rights, ...)
- versioning (origin, audit-trail, ...)

→ object network

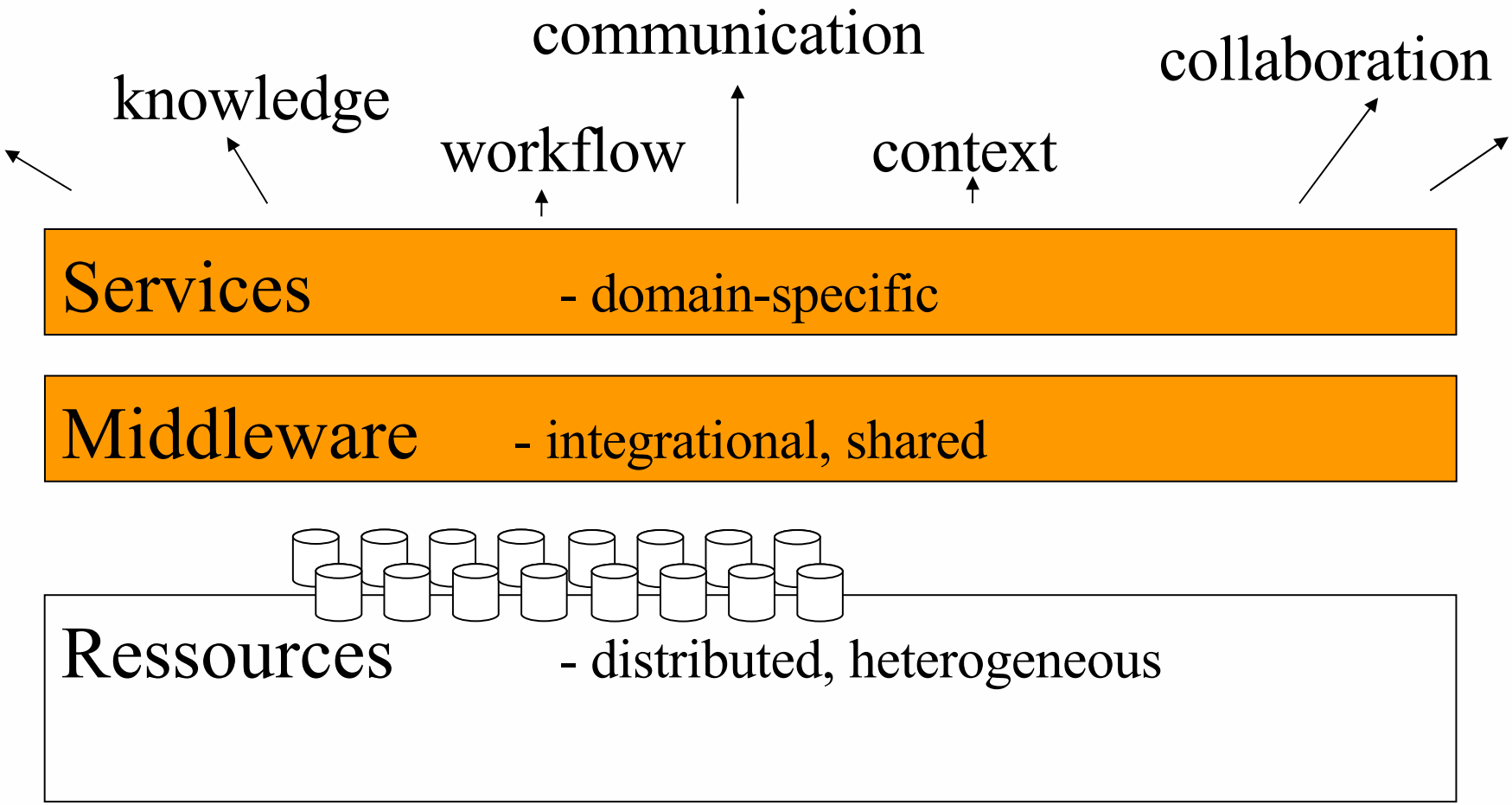
RDF triples – (subject, predicate, object)



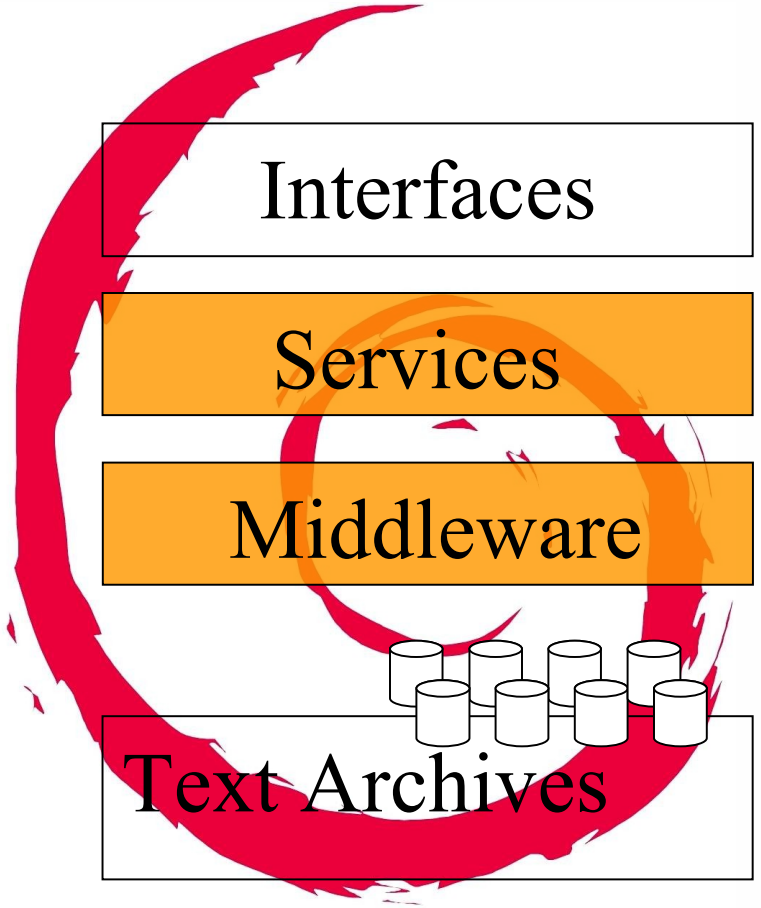
Overview

- TextGrid, Humanities and Grid Computing
- Objectives and requirements
- **Technology**

grid environment

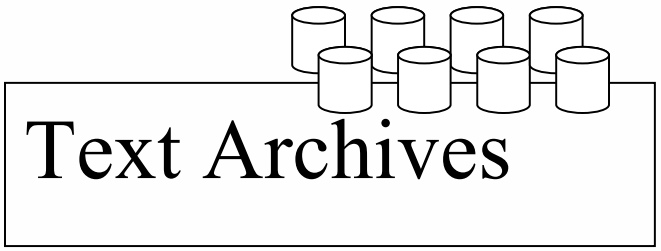


grid environment



Interfaces

- Graphical User Interfaces
  - Portal (GridSphere)
  - Rich Client (Eclipse)
- Providing a workbench
  - For configuring Services
  - specifying workflows
  - integrating interactive tools (editor, etc.)



- distributed, heterogeneous, autonomous data grid nodes
- integration (content, metadata)

$\forall \rightarrow$ interface to archives

$\forall \rightarrow$ policy

Services

- Proving encapsuled tools, e.g.
  - tokenizer, lemmatizer, dictionary services, streaming editor
  - workflow enacter
  - middleware interaction, e.g., registring and retrieving resources
- Based on standard (non-Grid) technology:
  - WSDL, SOAP
  - WSRF is handled in the middleware

Services = SOA

- Service Oriented Architecture
  - the new paradigm in IT
  - Web Services is the way to do SOA
- Mapping of organizational workflows
- SOA is infrastructure
  - Communication between processes
  - Registry of services
  - Orchestration of workflows
  - common security features

Middleware

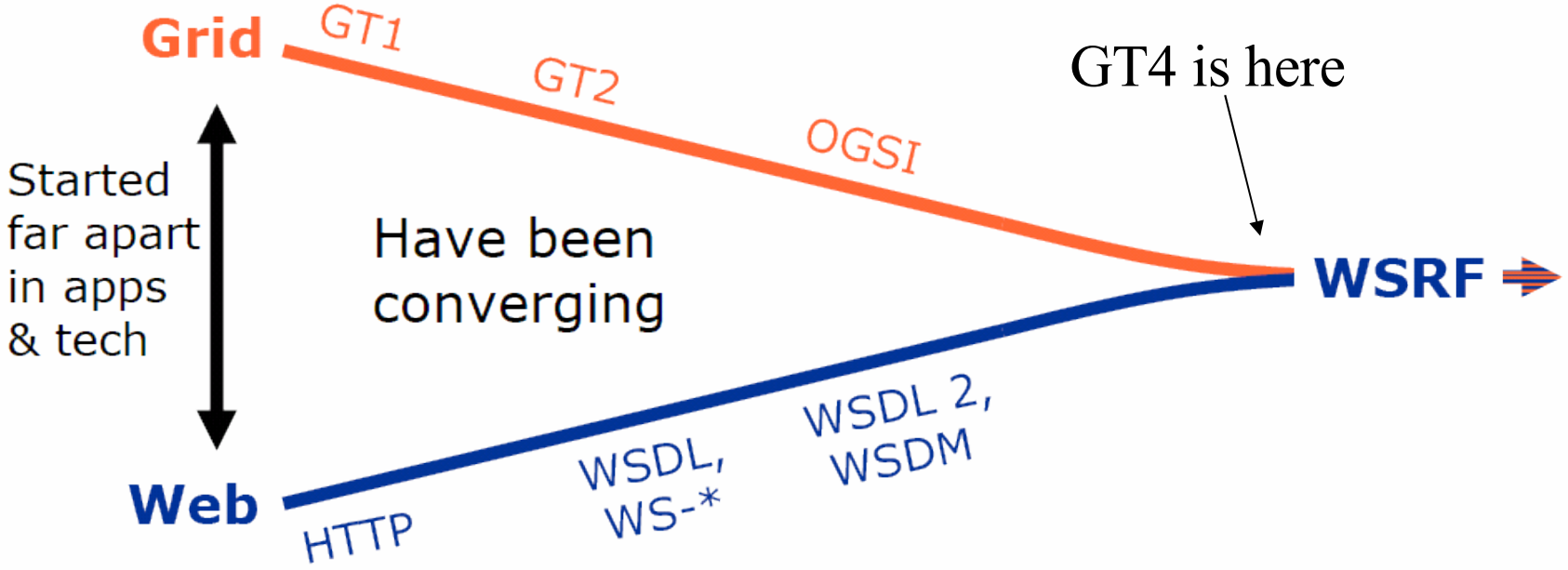
$\forall \rightarrow$ service layer

$\forall \rightarrow$ archives

$\forall \rightarrow$ D-Grid infrastructure

Middleware Globus Toolkit

- GGF/OGF-Standardization and Web Services development



GT4 and standards

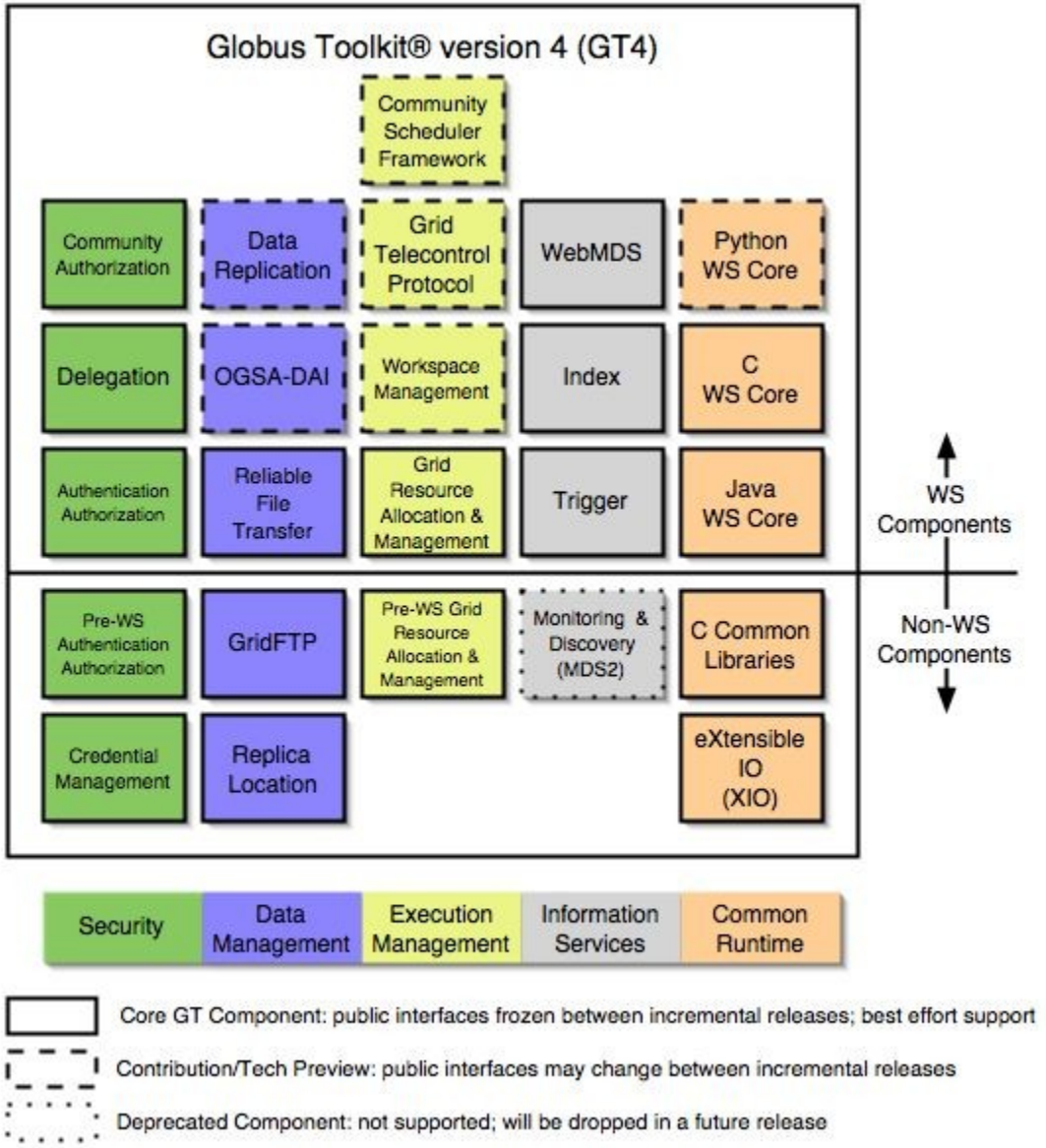
- SSL/TLS v1 (from OpenSSL) (IETF)
- X.509 Proxy Certificates (IETF)
- GridFTP v1.0 (OGF)
- WSRF and WS-N (OASIS)
- And others on the road to standardization:

  DAI, WS-Agreement, WSDL 2.0, WSDM, SAML, XACML

  http://www.mcs.anl.gov/~liming/primer/

Middleware GT4

http://www.globus.org/toolkit/about.html



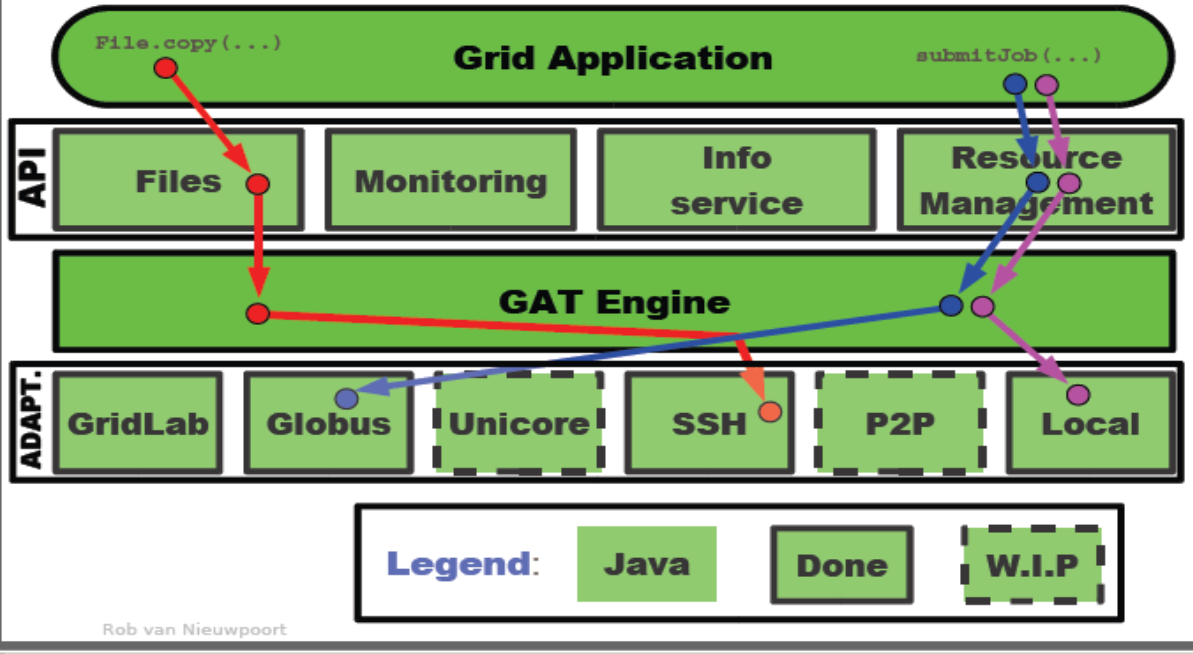
Grid Application Toolkit (GAT)

- GAT does not aim to replace existing “grid infrastructure.”
- GAT aims to provide a simple, clear interface to many different infrastructures
  - GRAM
  - Condor
  - Unicore
  - GridFTP
  - RFT…
- Open source, BSD-like licence

JavaGAT-tutorial.pdf (application/pdf Object) - Mozilla Firefox

http://www.cs.vu.nl/~rob/JavaGAT-tutorial.pdf

Java GAT Structure



Rob van Nieuwpoort

Saga

- The Simple API for Grid Applications
- OGF-Standard for GAT
- Lessons learned from GAT included
- Modularized
  - only SAGA core is finalized
  - https://forge.gridforum.org/sf/projects/saga-core-wg
- Implementations are on their way
- TextGrid plans to migrate to Saga

Middleware

Data Grid, Service Grid, Computational Grid

Information integration

→ static + dynamic meta/data, versioning

→ virtualisation, replica management

Security: AAI → GridShib (Shibboleth DFN-AAI)

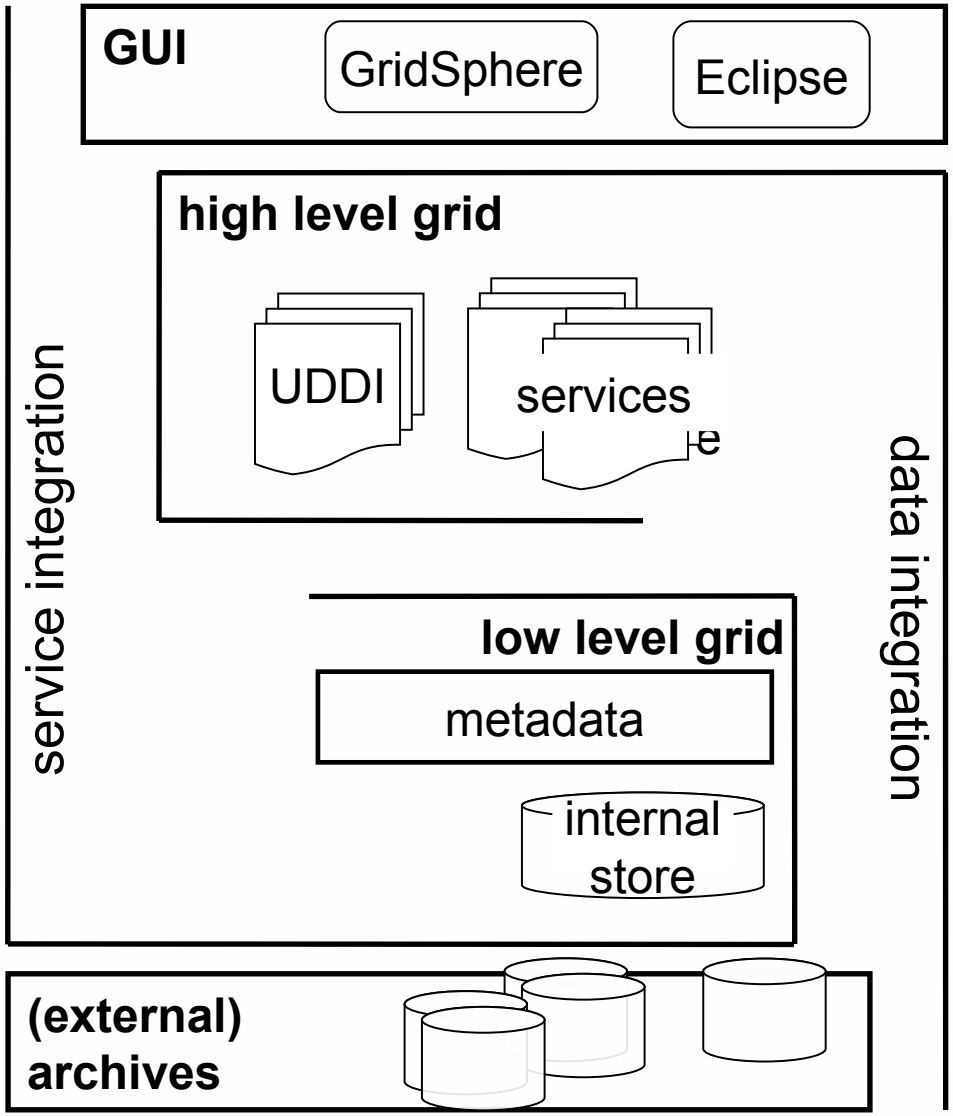
Workflow Management, Scheduling

Monitoring

License Management, SLA

…

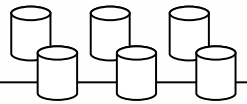
conclusions



Interfaces

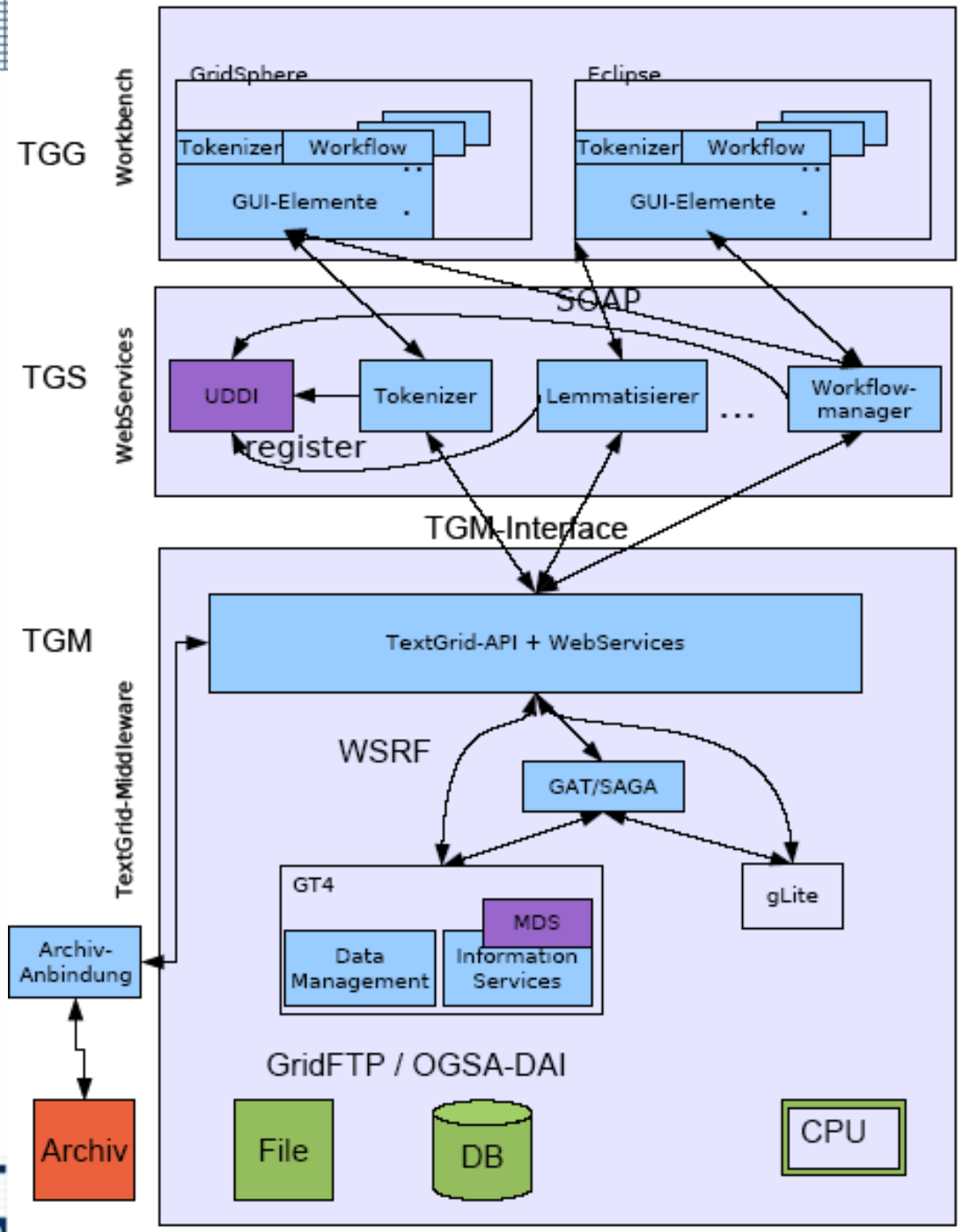
Services

Middleware



Archives

TextGrid architecture



Federal Ministry
of Education
and Research

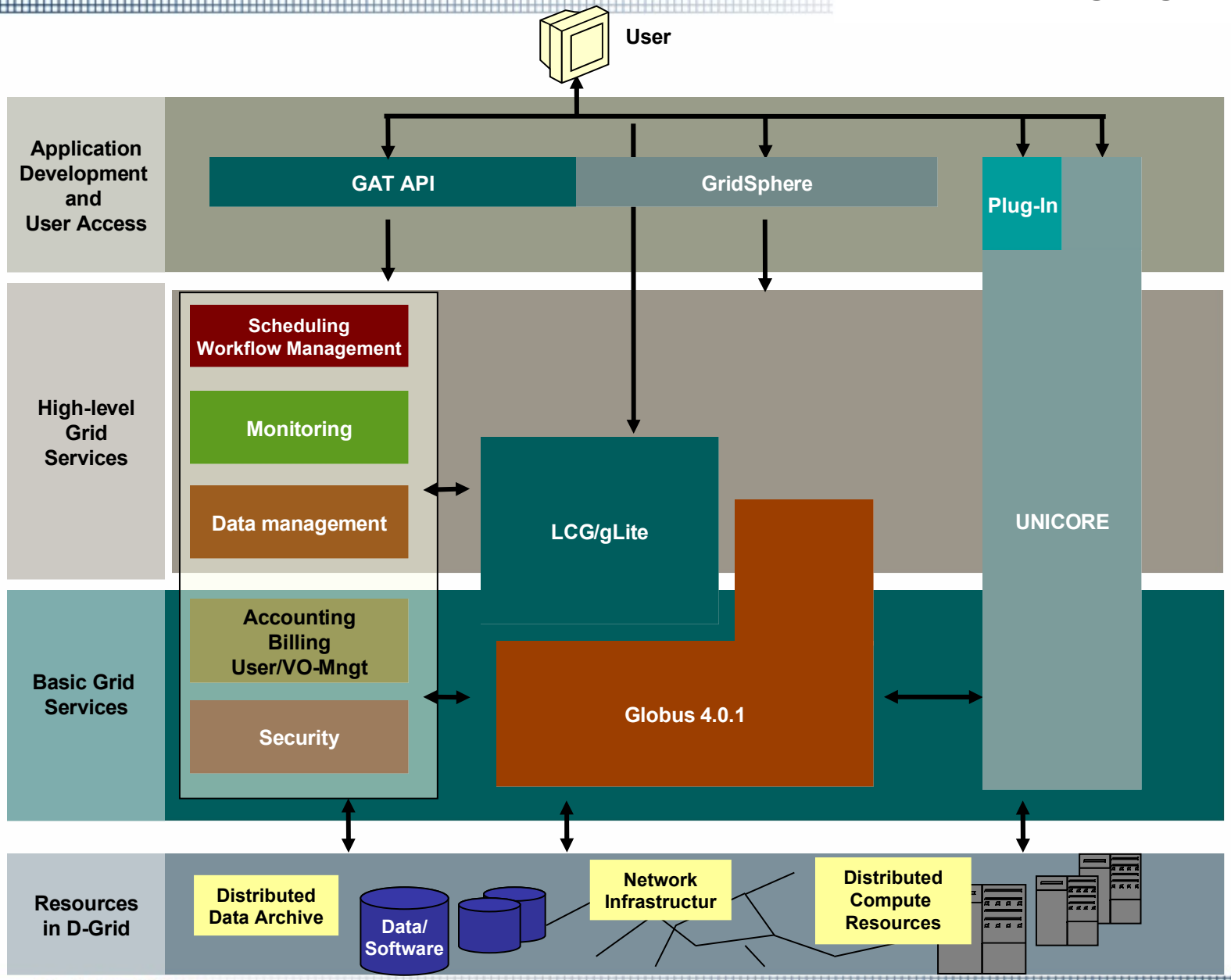
comments, questions, ... ?

thanks, for your attention !

TEXT GRID

DAASI International
Directory Applications for Advanced Security and Information Management

Peter Gietz
DAASI
peter (dot) gietz (at) daasi (dot) de


User
Application Development and User Access
GAT API
GridSphere
Plug-In
High-level Grid Services
Scheduling Workflow Management
Monitoring
Data management
LCG/gLite
UNICORE
Basic Grid Services
Accounting Billing User/VO-Mngt
Security
Globus 4.0.1
Resources in D-Grid
Distributed Data Archive
Data/ Software
Network Infrastructur
Distributed Compute Resources

~~non~~-hierarchic xml

<l> The IEEE (**<t>**eye **</l>**

<l> triple e**</t>** is first </l>

…

- annotation levels
  → typographic, structural, logic, semantic, text genetic, etc
- on-the-fly XML creation