MPG eScience Seminar 2008

Requirements of eScience and Grid Projects Towards Long-Term Preservation of Research Data

Jens Klump, GFZ Potsdam

Göttingen, 2008-06-19

Agenda

Results from the study „Requirements of eScience and Grid Projects Towards Long-Term Preservation of Research Data“.

- Context, dealing with research data
- Study outline, Materials, Methods
- Results from interviews
- Recommendations to stakeholders
- Conclusions

Context

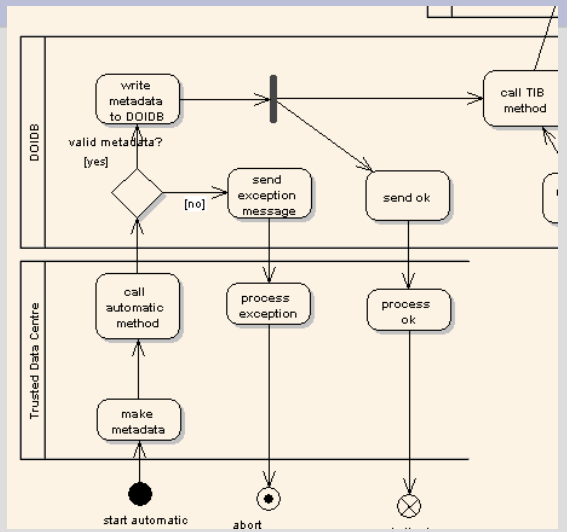
- Study commissioned by „Network of Expertise in Long-term STOrage of Digital Resources“ (nestor)
- Focus:
  - Application in the context of German projects,
  - Recommendations for digital long-term preservation in the context of eScience and Grid.
- General aspects of digital long-term preservation have already been covered by other studies.

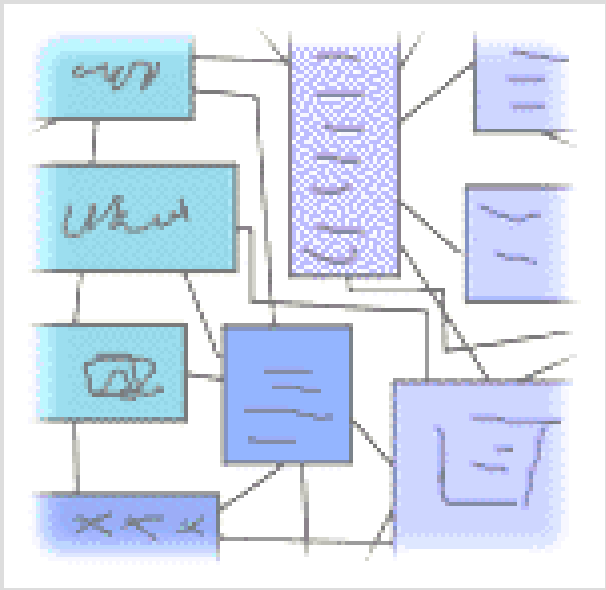
Why Archive Data?

- Archiving research data from projects ...
  - saves time and money by avoiding duplication,
  - improves quality of research by making results verifiable.
- Most data are at present not accessible.
- Archiving of research data is still unsystematic.

The Scientific Workflow

- Workflows in eBusiness and eGovernment are characterised by persistence and the requirement for transactional behaviour.
- Scientific workflows are characterised by *ad-hoc* changes, depending on the outcome of the preceding experimental step.





Definition of „Long-Term“

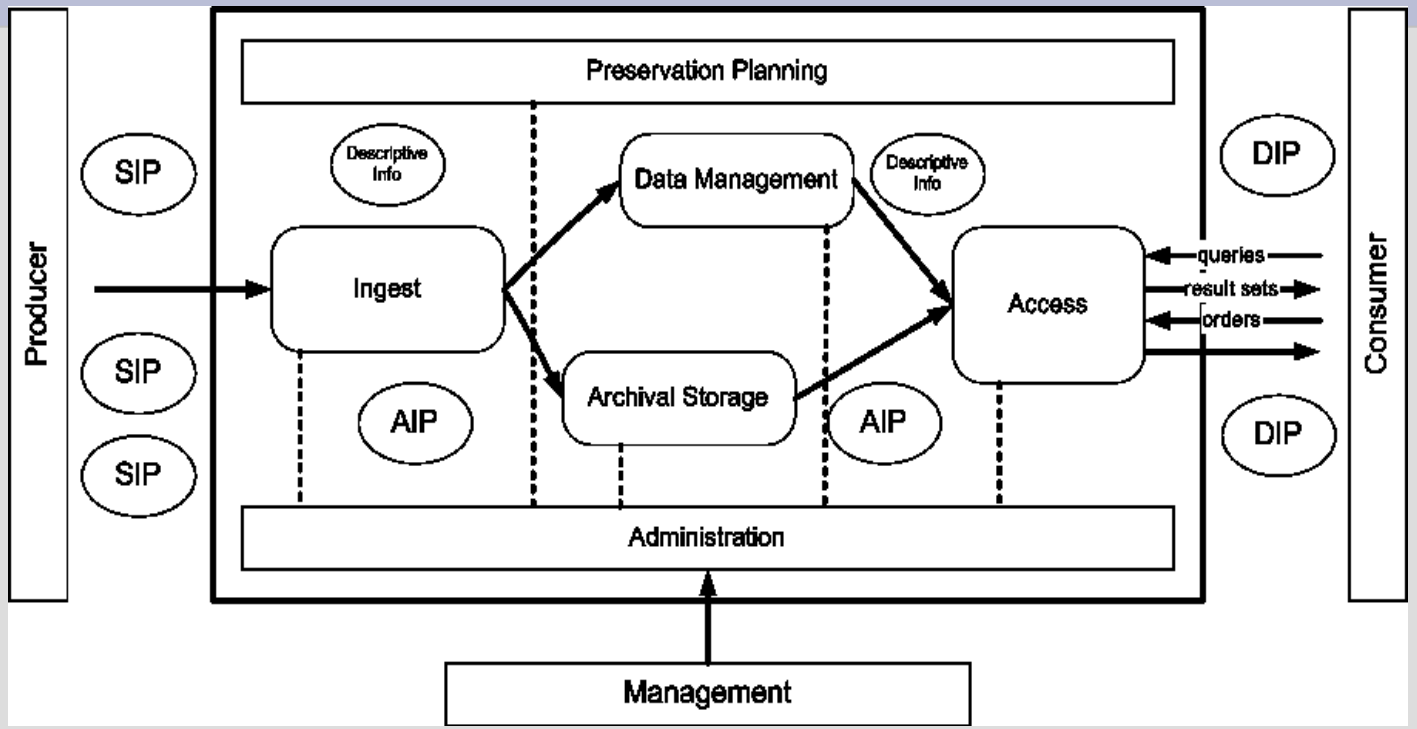
- Research data face their greatest risk of loss after the end of the funding period.
- For the context of this study we defined „long-term“ as „re-usable and reliable preservation well beyond the end of the project“.
- The duration of preservation is commonly defined in a policy or legal framework.

Definition of Long-Term

- Computer Science: > 5 yr.
- DFG, MPG: > 10 yr.
- Engineering: > 30 yr.
- Linguistics: > 100 yr.

- At present the domain of memory institutions (libraries, museums, archives).

OAIS as Reference Model



The OAIS reference model describes, how **humans and technical systems interact** in the long-term preservation of digital objects.

Materials and Methods

- Focus group: Projects funded by BMBF in April 2007.
- Method: Qualitative interviews with stakeholders in projects. The questionnaire served as a guideline for the the interviews.
- First results were presented at GES 2007 and discussed with stakeholders from projects.

Interview Partners

- Grid projects funded by BMBF and operating in April 2007
  - AstroGrid-D (astronomy, astrophysics)
  - C3-Grid (climate models)
  - HEP-Grid (high energy physics)
  - InGrid (engineering)
  - MediGrid (medical and life sciences)
  - TextGrid (arts and humanities text processing)

Interview Partners

- eScience projects funded by BMBF and operating in April 2007:
  - eSciDoc (scholarly communication platform)
  - Hyperimage (semantic image annotation)
  - Im Wissensnetz (information networking)
  - Ontoverse (life sciences)
  - SYNERGIE (computer sciences)
  - WIKINGER (information networking and repository)
  - WISENT (energy meteorology)

Interview Topics

- Data volume and complexity
- Dealing with metadata
- The semantic web
- Access to data and rights management
- Virtual organisations and sustainability
- Best practice examples
- Synergies between Grid/eScience and digital long-term preservation

Data Volumes and Complexity

- Après nous, le déluge (des donées)?
- The approach is very discipline specific.
- Data volumes vary by several orders of magnitude.
- Grid project data can be very complex.
- Policies range from 5 years (computer science) to indefinitely (linguistics).
- Operating costs of storage are a limiting factor.

Metadata

- Metadata standards are problematic.
- Standards are often not accepted in the communities and seen as under/over complex.
- Communities are still divided over comprehensive vs. light-weight metadata schemas.

Metadata

- Metadata on formats and file types are rarely collected. MIME-type is not enough!
- Too little attention is paid to file formats and their suitability for long-term archiving.
- Provenance and processing metadata vary.

Semantic Web

- Capture and processing of semantic relations between data objects is a key objective in many projects (predominantly eScience).
- Semantic relations to physical objects („internet of things“) are found in very specific contexts (e.g. pathology specimens).
- Capture and management of implicit knowledge is practiced in some projects.
- SME industrial partners are sceptical („processes can not yet be copied“).

Access and Rights

- The degree of data sharing depends on established practices in the communities.
- Machine readable licences would help.
- Distributed data storage would be interesting for SME, but access management is not yet of fine enough granularity.
- The role of the systems administrator is seen as critical (intransparent, lack of trust criteria).

Access and Rights

- Research is needed in identity and credentials management.
- Long-term management of certificates poses new questions (e.g. migration of keys, orphaned certificates).

VOs and Sustainability

- Only a minority of VO have policies on long-term preservation.
- Communities are aware of this issue and are formulating policies.
- In many cases, long-term preservation is seen as beyond the scope of the project.
- Roles and responsibilities in VO are often not formalised.
- More research is needed into the management of VO.

Recommendations

- Implementation of a testbed for long-term preservation in Grid environments.
- Research into application of Grid technologies for long-term preservation (e.g. format migration, ingest process, emulation, ...)
- Communication of best practice examples to communities.

Recommendations

- More effort into documentation of provenance, processes and implicit knowledge.
- Adaptation of semantic web technologies towards a semantic grid.
- Research into digital rights management (credentials, long-term aspects of certificates).
- Research into VO management.

Conclusions

- eScience and Grid projects are aware of the challenges of long-term preservation.
- A number of technical and organisational issues need to be resolved.
- Synergies between Grid technology and long-term preservation should be explored in a testbed project.
- More communication of best practices among and between communities is needed.

Acknowledgements

- Funding was provided by BMBF through the „Network of Expertise in Long-term STOrage of Digital Resources“ (nestor).
- The author would also like to thank the interview partners, the members of the nestor working group on Grid/eScience, and the participants at the GES 2007 nestor workshop for their input and discussions.