**4th e-Infrastructure Concertation**
Sophia Antipolis
5-6 December 2007

# Standardisation
# and
# e-Infrastructures

## DATA TRACK
## (Meeting Room IRIS 6 )

Chair : Krystyna Marek
Rapporteur: Wolfram Horstmann

# Session Objectives

- Find community specific requirements
- Find relations with standardization
- Find useful next steps

# Starting points

- Scientific data infrastructures: new field in FP7
  - Very different from networking and grids
    - Representing 15% non-IT disciplines from study sample
- Programme Objectives
  - New projects reinforcing Research Capacities
  - Develop ICT-based infrastructures
- Learning from communities
  - Standardization envisioned for later stages
- Relation to council conclusions
  - Stress publications and data

# Participants

- By call
  - Repository infrastructures
    - NMBD
    - DRIVER-II
    - EuroVO-DCA, EuroVO-AIDA
    - Genesi-DR
    - METAFOR
  - User communities
    - D4Science (Diligent)
  - Design studies
  - Scientific data infrastructures (2008)
    - Parse.insight, (Caspar) (preservation)
- Observers
  - D-Grid (Knowledge Layer)
  - BELIEF: for reporting

# Self-perception

- ‚Vertical‘ Communities
    - Complex, multidisciplinary disciplines with intreroperability challenges within the community
    - Specific, heterogeneous provenance of data
    - Users of standards (but also developers?)

# Metadata, Data and Formats

- Legacy: data collected for many years
  - proprietary encodings (e.g. vendor-driven)
- Bottom-up problem solving
  - instrument/methodology-driven
- Wide range of data volumes (GB vs. PB)
- Resources have complex life-cycle
  - Multidisciplinary: No common denominator
  - Formats differently applied across communities
    - Differentiation between metadata and data non always valid

# Focus on Interoperability

- Not the same as standards
- Problem oriented solutions
  - „Diversity of Formats not the main problem“
    - e.g. language harmonization by converter
    - e.g. running models again cheaper than reformat
  - Standardization as *a posteriori* process
    - „As opposed to industry“ (?)
  - Usage vs. Preservation
    - Actual requirements vs. sustainability

# Common standards usage

- Authentication and Authorization
- Authenticity of Resources
- Provenance information
  - Contextualize the creation situation
- Preservation

>> „But it‘s not our core business“

# A notion: ”Division of labour“

- Networks and GRIDs provide generic interop.
  - the research process is not immediately touched
- Preservation not done by the researcher
  - Responsibility of data-centers and data producers
- Curation / quality control
  - Collaboration with researchers needed
- Research process is community-driven

# Expectations on data interoperability / standards

- Access layer to a wide range of different resources needed
  - Not much horizontal data standardization
  - Only interface standardization
  - Virtualization of resources
  - Respecting (not developing) standards

# Standards-Use

| W3C | ISO | OASIS | IEEE | IETF | ETSI |
|---|---|---|---|---|---|
| • [all basics]<br>• Web Services (WSDL, SOAP)<br>•Ontologies/ Semantic Web (e.g. SKOS) | • Vocabularies (language, country, dates)<br>• Virtual research environments<br>• Geographic MetaData & information and services<br>• Archiving/OAIS | • Web Services (UDDI)<br>• A&A (SAML/ XACML)<br>• Business Markup (ebXML) | • Architecture (HLA)<br>• Simulation (DIS) | • No mention | • No mention |

Red = proactive contributons

# Standards-Use

| ***OGF*** | ***OAI*** | ***DCMI*** | ***LOC*** | ***IVOA*** *(subject based)* | ***Other*** *(subject based)* |
|---|---|---|---|---|---|
| • „Usage of other people‘s work“ | • Resource exposure/ aggregation (OAI-PMH)<br>• Object Re-use and Exchange (OAI-ORE) | • Simple Metadata (DCMES)<br>• Virtualizing (DC-Collection) | • Web-Service queries (SRU/W-CQL)<br>• … | • Metadata<br>• Resource Registry | • [Ontologies]<br>•… |

Red = proactive contributons

# Next steps

- Share lessons learnt in data-management
  - Simple forms of networking
  - Standards web-site
    - Functions / usage models
  - Forum
  - Mailing-list (?)
  - List of contact-persons
  - Bilateral discussions
  - Workshops

- Consultancy for generic standards

# Conclusions

- Research-process dominates data-management
- Distance from generic technology standards (e.g. networking/grids)
  - Cross-consultancy demand is acknowledged
- Heavy use of standards and even participation in standardization
- Further knowledge exchange appreciated

# Data infrastructure challenges

| **Function** | **Description** | **Example** |
|---|---|---|
| Services | Build service as interoperable entity | WS/SOA |
| Query | Find and access resources | SRU/W-CQL |
| Federation | Aggregate+normalize distributed resources | OAI-PMH/OAI-ORE |

# Data infrastructure challenges

*http://www.driver-support.eu/documents/DRIVER_Review_of_Technical_Standards.pdf*

| | | |
|---|---|---|
| Terminology services | Allow to interpret values of entity properties | Discovery and browsing through categorizations (e.g. ) |
| Registration services | To assign (persistent) identifiers to resources. | „Stay when URLs change…“ (e.g. DOI, URN) / identify conceptual/non-dig resources |
| Resolution services | To locate resources, typically from an identifier | „Content negotiation“ (e.g. find the right version of an image) |
| Authentication/Authorization services | To allow user specific environments | Integrated user management (e.g. SAML, XACML; Shibboleth) |
| Text mining and content processing | To allow [automatic] entity recognition and processing. | Citation analysis from the full text of ePrints. Relate to data (e.g. proteins …) |
| [Meta]data registries | To allow recording and relating data models. | … (e.g. XSchemas or RDFSchemas) |
| Service & coll-ection registries | To relate repositories content to services that can use that content. | DCMI Collection Description Application Profile and Service Description formats (e.g. WSDL ) |