

Breakout Group

# Content Aggregation and Networking Architectures

*OAI5 – CERN, Geneva 19-APR-2007*

Susanna Mornati and Wolfram Horstmann

# *Participants*

ALLINSON, Julie; UKOLN, University of Bath  
BADOLATO, Anne-Marie; INIST-CNRS  
CALLAN, Paula; Queensland University of Technology  
CEREJO, Clyde; University of Zurich  
EL KASMI, Hicham; Université Libre de Bruxelles  
ERLANDSEN, Morten; University of Oslo  
HAGEDORN, Kat; University of Michigan  
HEY, Jessie; University of Southampton;  
HOLMQVIST, Kristoffer; Lund University Libraries, Head Office  
HUNTER, Philip; Edinburgh University Library  
KALOYANOVA, Stefka; FAO  
KRICHEL, Thomas; Long Island University  
MANGHI, Paolo; Consiglio Nazionale delle Ricerche (CNR)  
MICHELOUD, Marylene; RERO - Library Network of Western Switzerland  
NOVAK, Petr; State Technical Library  
PARKOLA, Tomasz; Poznan Supercomputing and Networking Center  
ROBERTSON, R. John; University of Strathclyde  
RUIJGROK, Peter; Utrecht University Library  
SACCHI, Simone; CIB - University of Bologna  
SMITH, MacKenzie; MIT Libraries  
VAN DE SOMPEL, Herbert; Los Alamos National Laboratory  
VAN LUIJT, Martin; Utrecht University Library  
WARNER, Simeon; Cornell Information Science

# *Overview*

- After two brief presentations on current issues and challenges in content aggregations and networking architectures made by the breakout group organizers an open and intense discussion started.
- Topics related to content aggregation raised most interest.

# *Main Topics discussed*

- OAster Wishlist
- Ambiguous Identifiers
- Graph-based representations of resources
- Detecting duplicates
- Classification, particularly subject-based
- Full-text availability
- Author identification
- Rights expressions

# *OAIster Wishlist*

The current state in content aggregation as summarized in the form of a wishlist what a major service provider like OAIster expects from local repositories:

- good metadata
- separate identifiers (e.g. record + resource)
- subject metadata
- rights information

# *Ambiguous Identifiers*

- Clear differentiation between identifiers of records vs. (various) resources, versions etc. is required
- But who or what provides Identifiers: machine based only or humans needed?
- Mandatory validation of OAI identifier concept in each repository implementation would be most helpful
- Universal resolving mechanism for DOI, URN etc. for resource identifiers needed?
- Proposed solution: focus on object representation identified by URI (e.g. ORE)

# *Graph based representations*

- Even the most simple resource representation as metadata+resource is a compound object
- Machine readable graphs that can be harvested are needed
- Created at the level of the data provider
- Considered necessary to disambiguate resources
- Important to create a „overlay information space“ made of identifiers
- ePrints AP – as possible specific model in the overlay environment

# *Detecting duplicates*

- Duplicates still a problem
  - Rough quantitative guess in OAIster:  
 $1000 > x > \ln$
- Software solutions in place (but laborious?)
- Grouping vs. deduplicating
  - duplicates might actually be different versions
  - duplicates serve as secure copies



# *Classification*

- A universal scheme is not feasible
- Still: better solutions needed, enabling advanced browsing and search
- Text-mining based approaches are promising (e.g. OAISTER project)
- Human-based approaches also work (aided by learning algorithms, e.g. REPEC)

# *Full-text availability*

- Large amounts of metadata-only records „spoil“ the aggregations
- Differentiation between record types (metadata only vs. resource) needed
- Most efficiently achieved on the level of repository software
- Possible solution through unambiguous link representation (e.g. ORE, see above)

# *Author identification*

- Poses problems, partly solvable
- Exploiting national authority files possible
- But changes and multiple positions not represented
- Dynamic relationships represented in appropriate object models (e.g. ePrints AP)
- Identify people by URIs

# *Rights expressions*

- Rights information frequently missing
- Various schemes for rights applied
- Compound objects pose a „rights interoperability“ problem
- CC / RDF snippet an optimal solution