INTEREST - INTERoperation for Exploitation, Science and Technology

Keith G. Jeffery, STFC; Council Rutherford Appleton Laboratory, United Kingdom
Anne Asserson, University of Bergen, Research Department, Norway

Abstract
This paper addresses the topic of interoperation of Grey resources. The title should be read as INTERoperation for Exploitation, Science and Technology. It builds on work by the authors published in previous GL conferences. The method is architectural analysis and comparison. The costs of the study are negligible, but of course the costs of implementing any solution are considerable. The result/conclusion is that CERIF (Common European Research Information Format) is the essential component to meet the requirements and is applicable – to a greater or lesser degree - in all architectural solutions.
Our GL9 (2007) paper proposed a Grey landscape architecture and identified the need for (1) excellent metadata (to improve discovery and control usage), (2) an institutional document repository of (or including) grey, (3) an institutional CRIS (Current Research Information System) for the contextual research information, (4) linkage between the document repository and the CRIS of an institution and thence (in a controlled manner with formal descriptive and restrictive metadata) to other institutions, (5) an e-research repository of research datasets and software, (6) linkage between the e-research repository and the CRIS of an institution and thence (in a controlled manner with formal descriptive and restrictive metadata) to other institutions, (7) an institutional policy to mandate deposition of the material with appropriate metadata.
These very requirements define the components for interoperation of Grey resources, and their interoperation with other resources to provide a holistic support for R&D. Indeed they can be extended (via the CRIS) to interoperation with other management systems of an organisation such as finance, human resources, project management, production control etc.
However, the capability for interoperation can be provided in several implemented architectures. This paper discusses the advantages and disadvantages of different solutions including experience of their use. This analysis and experience is then applied to the grey environment. Remote and local wrapping of resources, cataloguing techniques and a full compliant model are discussed as well as harvesting technology. It concludes that – particularly for the grey environment – the optimal architecture involves formal syntax (structure of information) and defined semantics (meaning of information) as defined by CERIF.

Background

The Basis of the Paper

(1) Grey Literature repositories are much improved for the end-user (in integrity, relevance, quality and utility) when linked with a CERIF-CRIS or a CRIS capable of interoperating with other CRIS using CERIF;
(2) Although it is possible to link Grey literature repositories independently of CRIS if they store - for example, Dublin Core (DC) metadata and use OAI-PMH protocol for interoperation and OAISTER for searching – we propose that interlinking CERIF-CRIS (or CRIS capable of interoperating using CERIF) is better because of the formal syntax and declared semantics: CERIF contains sufficient metadata (Jeffery 2000) to provide better recall and relevance in retrieval than OAI-PMH-linked DC metadata-based systems;
At GL9 (Jeffery and Asserson 2007) we presented an architecture for utilising CRIS interlinked tightly using CERIF providing access to Grey literature repositories (both publications and research datasets/software). This paper attempts to present and compare architectures utilising CRIS (with assumed linked local grey literature institutional repositories) for interoperation. Of course (and presented in section 8) the ideal architecture is the use of a native CERIF-CRIS.

Metadata

Metadata may be classified into kinds: schema, navigational, associative with the latter partitioned into descriptive, restrictive and supportive (Jeffery 2000). Apart from harvesting, which ignores the syntax (structure) and semantics (meaning) of the data and just does text string searching, all architectures rely on a predicate query over a known schema (available or derived by schema reconciliation) thus allowing search terms or values to be related to an entity/attribute and thus to a domain. Example: the string 'green' could occur under attribute 'family name' in entity 'person' or within attribute 'abstract' or 'title' in entity 'project'. The use of query under a schema ensures that the query is meaningful and should have adequate recall (coverage) and relevance (precision).
Most techniques rely on navigational metadata to access hosts of CRISs. The catalog techniques use in addition associative descriptive metadata to perform the first pass search – rather like the harvesting