

Grey in the R&D Process

Keith G. Jeffery

CCLRC Rutherford Appleton Laboratory, United Kingdom

Anne Asserson

University of Bergen, UiB, Norway

Abstract

The rate of acquisition of data, its structuring into information and its interpretation as knowledge is increasing rapidly. There are more active researchers now than ever and the output of white publications per researcher is increasing. The output of grey publications is orders of magnitude greater. Past techniques of experts (librarians) cataloguing manually with metadata each publication do not scale. The problem is to find ways to manage this resource.

The hypothesis is in 4 parts: (a) that the R&D process itself provides some context for managing the information; (b) that linking the records of the process to the publications provides this context; (c) that questions of curation and provenance are addressed automatically in such an environment; (d) that such an environment integrates grey and white literature and other R&D outputs such as software, data, products and patents.

At UiB the emphasis of the work has been on assessment of the research output - especially publications - linked in context with records of the researchers, their organisational units, and related CRIS (Current Research Information System) information (the FRIDA system which is mostly CERIF-compatible).

At CCLRC the emphasis of the work has been on the production of an open access repository of publication outputs from the organisation (ePubs), linked to the CERIF-compatible CDR (Corporate Data Repository) CRIS and thus to other research outputs with associated metadata.

The recording of the data provides the context including the workflow of the R&D process, history and provenance. Grey documents produced as early ideas are captured in a temporal and organisational context, just as well as white publications, via the linked repository. CERIF allows, in a multidimensional framework, deduction or induction of relationships between documents – for example between a grey internal report and a white published paper - and with other research outputs. Furthermore, relationships between documents can be expressed explicitly: references and / or citations can be recorded. In this way a rich context for understanding the R&D output is provided, including versions, history and provenance.

Recording facts once in a structured R&D process environment and then re-using them in many ways reduces - by automated provision assistance - the need for user input of metadata to describe research outputs (especially grey literature) and thus addresses the scalability problem.

The costs including all staff, overheads, equipment, software etc for CCLRC are as follows:

(a) Development of the CDR (Corporate Data Repository CERIF-compatible): (i) Pilot phase: 30kEUR; (ii) Production phase (~ 1 year): 80kEUR; (iii) Annual maintenance including integration with the ePubs system and with our workflow environment: 30kEUR

(b) Development of the ePubs Open Access Institutional Repository: (i) 80kEUR (~ 1 year); (ii) Annual maintenance including integration with CDR 50kEUR

We are now (with integrating CDR, ePubs and workflow) developing re-engineered business processes: this programme of work is estimated at ~ 300kEUR per year for 2 years and relies on CDR and ePubs. We believe the benefits in improved effectiveness and efficiency will run at ~ 4mEUR per year.

The costs for the University of Bergen are as follows: The total cost of developing our national system Frida is 2.250.000 NOK for the 4 universities and the cost of UiB (only for developing) is approx 500.000 NOK.