Controlled access to data in EPrints

1. Introduction

The RoaDMaP Project emailed two discussion lists in May 2013 to see if others were interested in fine-grained access management to content in an EPrints repository. We are piloting a research data repository for the University of Leeds; we also use EPrints for our Digital Library service. We anticipate both services will contain (or reference) materials which can only be made available to particular groups of users e.g. for ethical reasons.

Many responses were received on and off list. The main points are summarised in Section 3 below.

2. Email sent to lists 8th May 2013

JISCMRD@JISCMAIL.AC.UK and eprints-tech@ecs.soton.ac.uk

Dear colleagues

We are looking at how to manage temporary or longer term restricted access to data within an institutional research data repository. Our starting point is that, where possible, data should be made openly available with as few restrictions as possible. However, we're aware that there are various reasons why it may not be possible - or appropriate - to share data openly.

We are piloting EPrints for research data. We need the means to create different access levels within EPrints which can then be assigned to users.

Broadly, we're looking at an access model where research data can be:

- *Public - openly available without any registration requirement*
- *Registered - the user is required to register to access content; this could be linked to authentication protocols such as Shibboleth. Access may be time limited.*
- *Approved – the potential user makes a case to use the data; approval relates to specified data within the repository. Access may be time limited.*
- *Embargoed – metadata and data held in the dark for a specific period*

We are interested in whether other institutions are planning - or have put in practice - granular access to research data and what this looks like.

We would also like to feed back requirements to Southampton EPrints Services; if we can scope requirements before July, Leeds has funds to put towards development costs.

Is anyone else interested in this functionality?

Thanks for any comments

Best wishes
Rachel

***

Rachel Proudfoot
Project Manager
RoaDMaP: Leeds Research Data Management Pilot
http://library.leeds.ac.uk/roadmap-project
Tel: 0113 343 4554

3. Main points and questions arising from the responses

3.1 Level of interest

- Several institutions expressed an interest in more granular control of access to EPrints content (University of Southampton, University of East London, University of the West of England, Glasgow School of Art, University of Hertfordshire & Hungarian Academy of Sciences). For example, GSA noted some of their data is commercially and ethically sensitive.

3.2 Existing functionality in EPrints

- Some access scenarios are supported 'out of the box' through EPrints embargo and request button features. However, these may not be sufficient for all access scenarios: for example, time limited access.

3.3 Access control: how, why, pros and cons

- There was some support for using Shibboleth as a means to make data available to the research community.
- For scalability, access control should require as little manual intervention as possible.
- There were differences in opinion about the pros and cons of offering 'Registered access' to data - for example, if registration is offered, will depositors tend to request it even where the data could be made openly available? To what extent is a registration requirement a barrier for potential re-users?
- Is the primary use case for registration to restrict access to sensitive data, or is it to monitor usage of data for statistical reporting and in case of breach of use conditions? As Chris Rusbridge noted, even "anonymised" data may be disclosive if combined with other data, suggesting the need for registration and agreement to terms for some data.
- Who makes the decision whether to grant access to a sensitive dataset if the PI/data

steward / data owner has left the institution?
- Parallels were drawn with other systems: for example, management of neutron spallation data by ISIS http://www.isis.stfc.ac.uk/groups/computing/data/icat11680.html . This service doesn't use EPrints but provides granular access to data - raw data + metadata is typically private to PI and nominees for 3 year, occasionally longer; some data is open; access to data results is determined by the PI; some data is private forever. Access to most data requires registration.
- How will access to very sensitive data be managed: what security controls are needed and, if data is *that* sensitive, should it be on a public server at all?

3.4 Licensing

- As well as controlled access, licence and re-use conditions should be considered. Some commentators questioned whether the CC0 licence is appropriate for data; others highlighted that incompatible licences with different re-use conditions will make it difficult or impossible to combine data sets.

3.5 What is 'open' data?

- Is a registration requirement compatible with 'open' data?
- Terminology is important. We need a name for data which is available for free upon registration; there was some feeling this should *not* include the word 'open'. Suggestions included 'transparent', 'managed', 'controlled' and 'registered'.

3.6. System architecture

- Would a Research Information System - such as Symplectic - form part of the deposit workflow for data?
- Will institutions use the same instance of EPrints for all types of content (papers, theses, data) or have a separate instance for research data?
- The access control workflow also depends on where the data is physically located. For example, we will be trialling archival storage offered by Arkivum exploring how to serve large and/or sensitive data sets. Commentary from Bill Worthington and Matthew Addis suggested at least two workflows:
  (i) large data sets located in archival storage: request-restore-ready-deliver
  (ii) sensitive data sets: request-decide-retrieve(or decline)-deliver

There was also feedback that it would be better to develop any new functionality collaboratively with the EPrints community rather than specifically for the University of Leeds; we agree.

4. Conclusions (RP in personal capacity)

- Monitoring usage is not a sufficient use case to require registration; where feasible, metadata and research data should be openly available with as few restrictions as

possible to avoid licence clashes.

- Registration could be offered as an optional extra - if there are benefits to the registrants (e.g. alerts of new / updated data sets).
- A registration function is required where research data has been obtained under agreements promising data access and re-use control; it would be difficult to change these agreements retrospectively but it may be possible to apply a rule based processes to semi-automate access to the data e.g. automatic access to those with an academic domain email address.
- Although we can encourage maximum openness as best practice (for data without commercial or ethical requirements for restriction), research data deposit is new in several subject disciplines and some level of control may be the price we pay to populate data repositories during a period of cultural change.
- Data on request - requiring an access decision by someone - is labour intensive and potentially stores issues for the future as data owners move on; however, there is likely to be an ongoing requirement for this level of control. Even very open data services such as the EMBL-EBI have some data which is only accessible by application to a committee. Further work is needed to map out these request workflows to understand the role of an institutional data repository in the request process and the delivery mechanism(s) for the data.

5. What next?

We are scoping requirements to feed back to Southampton with a view to developing additional functionality which can be offered back to the EPrints community. We will make the requirements available for comment shortly.

Rachel Proudfoot on behalf of the RoaDMaP Project
http://library.leeds.ac.uk/roadmap-project
18 June 2013