## Project Information

<table>
<thead>
<tr>
<th>Project Identifier</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Title</td>
<td>Repository Junction Broker (RJ Broker)</td>
</tr>
<tr>
<td>Project Hashtag</td>
<td></td>
</tr>
<tr>
<td>Start Date</td>
<td>1st April 2012</td>
</tr>
<tr>
<td>End Date</td>
<td>31st March 2013 extended to 31st July 2013</td>
</tr>
<tr>
<td>Lead Institution</td>
<td>EDINA</td>
</tr>
<tr>
<td>Project Director</td>
<td>Christine Rees</td>
</tr>
<tr>
<td>Project Manager</td>
<td>Muriel Mewissen</td>
</tr>
<tr>
<td>Contact email</td>
<td><a href="mailto:Muriel.Mewissen@ed.ac.uk">Muriel.Mewissen@ed.ac.uk</a></td>
</tr>
<tr>
<td>Partner Institutions</td>
<td></td>
</tr>
<tr>
<td>Project Web URL</td>
<td>Project Page: <a href="http://edina.ac.uk/projects/RJB_summary.html">http://edina.ac.uk/projects/RJB_summary.html</a></td>
</tr>
<tr>
<td>Programme Name</td>
<td>Repositories and Curation Shared Infrastructure Programme</td>
</tr>
<tr>
<td>Programme Manager</td>
<td>Balviar Notay</td>
</tr>
</tbody>
</table>

## Document Information

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Muriel Mewissen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Role(s)</td>
<td>Project Manager</td>
</tr>
<tr>
<td>Date</td>
<td>24 June 2013</td>
</tr>
<tr>
<td>Filename</td>
<td></td>
</tr>
<tr>
<td>URL</td>
<td></td>
</tr>
<tr>
<td>Access</td>
<td>This report is for general dissemination</td>
</tr>
</tbody>
</table>

## Document History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>24 June 2013</td>
<td>First draft</td>
</tr>
<tr>
<td>1.0</td>
<td>29 September 2013</td>
<td>First external version</td>
</tr>
</tbody>
</table>
# Table of Contents

**JISC**

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ACKNOWLEDGEMENTS</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>PROJECT SUMMARY</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>MAIN BODY OF REPORT</td>
<td>3</td>
</tr>
<tr>
<td>3.1</td>
<td>PROJECT OUTPUTS AND OUTCOMES</td>
<td>3</td>
</tr>
<tr>
<td>3.2</td>
<td>HOW DID YOU GO ABOUT ACHIEVING YOUR OUTPUTS / OUTCOMES?</td>
<td>4</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Project Management, Dissemination and Outreach</td>
<td>4</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Data Gathering</td>
<td>6</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Software Development, Essential Functionality &amp; Additional Functionality</td>
<td>7</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Data Trials, Use Cases, Scalability &amp; Early Adopters</td>
<td>10</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Service Environment Setup &amp; Sustainability</td>
<td>13</td>
</tr>
<tr>
<td>3.3</td>
<td>LESSONS LEARNED</td>
<td>14</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Fragmentation of IR Landscape</td>
<td>14</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Support for Rich Metadata in DSpace</td>
<td>14</td>
</tr>
<tr>
<td>3.3.3</td>
<td>EPrints SWORD 2.0 Implementation</td>
<td>15</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Full Text</td>
<td>15</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Changing Landscape</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>CONCLUSIONS</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>REFERENCES</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>ANNEXES</td>
<td>17</td>
</tr>
<tr>
<td>6.1</td>
<td>ANNEXE 1: LOGIC EVALUATION PLANNER</td>
<td>17</td>
</tr>
<tr>
<td>6.2</td>
<td>ANNEXE 2: UK REPOSITORYNET+ GENERIC EVALUATION QUESTION</td>
<td>17</td>
</tr>
<tr>
<td>6.3</td>
<td>ANNEXE 3: USER FEEDBACK FORMS</td>
<td>17</td>
</tr>
<tr>
<td>6.3.1</td>
<td>For Data Providers</td>
<td>17</td>
</tr>
<tr>
<td>6.3.2</td>
<td>For Institutional Repositories</td>
<td>17</td>
</tr>
<tr>
<td>6.4</td>
<td>ANNEXE 4: LEGAL AGREEMENT TEMPLATES</td>
<td>17</td>
</tr>
<tr>
<td>6.4.1</td>
<td>For Data Providers</td>
<td>17</td>
</tr>
<tr>
<td>6.4.2</td>
<td>For Institutional Repositories</td>
<td>17</td>
</tr>
<tr>
<td>6.5</td>
<td>ANNEXE 5: REVIEW OF METADATA STANDARDS FOR POTENTIAL SUPPORT BY THE RJ BROKER</td>
<td>17</td>
</tr>
<tr>
<td>6.6</td>
<td>ANNEXE 6: PROPOSED SERVICE OFFERING</td>
<td>17</td>
</tr>
<tr>
<td>6.6.1</td>
<td>ORI</td>
<td>17</td>
</tr>
<tr>
<td>6.6.2</td>
<td>RJ Broker</td>
<td>17</td>
</tr>
<tr>
<td>6.7</td>
<td>ANNEXE 7: PROPOSED SERVICE LEVEL DESCRIPTION</td>
<td>17</td>
</tr>
<tr>
<td>6.7.1</td>
<td>ORI</td>
<td>17</td>
</tr>
<tr>
<td>6.7.2</td>
<td>RJ Broker</td>
<td>17</td>
</tr>
<tr>
<td>6.8</td>
<td>ANNEXE 8: PROPOSED SERVICE ROADMAP</td>
<td>17</td>
</tr>
<tr>
<td>6.8.1</td>
<td>ORI</td>
<td>17</td>
</tr>
<tr>
<td>6.8.2</td>
<td>RJ Broker</td>
<td>17</td>
</tr>
<tr>
<td>6.9</td>
<td>ANNEXE 9: HELPDESK WORKFLOW</td>
<td>17</td>
</tr>
</tbody>
</table>
1 Acknowledgements

This project was funded by JISC and undertaken as part of the Repositories and Curation Shared Infrastructure Programme. The team would like to thank the UK RepositoryNet+ (RepNet) (1) for their support. We would also like to acknowledge the contributions of projects and groups who supplied us with data and took part in the trials to transfer deposits from the Repository Junction Broker (RJ Broker) to repositories.

2 Project Summary

The Repository Junction Broker (RJ Broker) project carried out the design, development and testing of a standalone middleware tool for brokering the delivery of research output between multiple data suppliers such as publishers and subject repositories to multiple institutional repositories (IRs). The RJ Broker accepts data objects, parses the metadata to determine appropriate target repositories and transfers the data to registered repositories. A web based user interface and APIs are also available for browsing and downloading and allow direct access to the unrestricted content of the RJ Broker. In addition, notification functionality is in place to provide regular emails to IRs that have yet to register in order to alert them of the presence of relevant content for their IR in the RJ Broker.

The project put in place the infrastructure required to offer the RJ Broker as a delivery service for research output. This includes the software application, hardware environment (with installations for development, testing and production), proposed legal agreement templates for embargoed content, Service Level Definition (SLD), development roadmap and technical and user documentation.

The RJ Broker collects records in an EPrints repository, returning a unique identifier to the data supplier. Affiliation information is extracted to identify appropriate target repositories using EDINA Organisation and Repository Identification (ORI) middleware tool. Open access records are transferred to all registered target repositories using SWORD. Those with a publisher embargo are only transferred to IRs that signed the RJ Broker legal agreement on compliance to embargo periods.

Successful data transfer trials between data suppliers, the broker and data consumers were carried out. Records were received from Nature Publishing Group and Europe PubMed Central. Deposits were made to DSpace repositories at the University of Edinburgh and Imperial College London; and EPrints repositories at City University London and OpenDepot.org. Work is on-going at St Andrews University and Massachusetts Institute of Technology.

The RJ Broker supports Open Access and the objectives of the Open Access Implementation Group of increasing the number of deposits to UK repositories by minimising the efforts required of potential depositors and repository staff, and thereby maximises distribution and exposure of research outputs.

3 Main Body of Report

3.1 Project Outputs and Outcomes

<table>
<thead>
<tr>
<th>Output / Outcome Type</th>
<th>Brief Description and URLs (where applicable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Project plans</td>
<td>A project plan for the first year was produced and delivered to JISC and RepNet on the 28th May 2012. A separate work plan for the four months extension was produced and the final version agreed on the 17th June 2013.</td>
</tr>
<tr>
<td>2. Final report</td>
<td>This document to be handed in to JISC and RepNet.</td>
</tr>
<tr>
<td>3. RJ Broker middleware tool</td>
<td>The RJ Broker tool comprises all the software needed to install, test and run the application. In particular, it includes the database, APIs, web based GUI, notification script and the bespoke data importers scripts developed for the data suppliers and consumers involved in the trials.</td>
</tr>
<tr>
<td>4. Service infrastructure</td>
<td>Development, test and production environments including backup and redundancy have been set up at EDINA and are ready for future service deployment. The development broker is available at <a href="http://devel.edina.ac.uk:1203/">http://devel.edina.ac.uk:1203/</a> The test broker is available at <a href="http://ribqalb.edina.ac.uk/">http://ribqalb.edina.ac.uk/</a></td>
</tr>
</tbody>
</table>
5. Technical documentation

User guide, technical manual and FAQs documents are available online in the “About” section of the broker website, [http://broker.edina.ac.uk/information.html](http://broker.edina.ac.uk/information.html).

6. Template legal agreements

Template agreements for receiving embargo records between a data supplier and RJ Broker and between RJ Broker and a data consumer, see Annexe 4: Legal Agreement Templates.

7. Metadata standard review

Review of the metadata standard relevant to the RJ Broker, with a short report provided to JISC and RepNet on the 18th January 2013, see Annexe 5: Review of Metadata Standards for Potential Support by the RJ Broker.

8. Use cases

Data trials were set up to develop and test the RJ Broker. Nature Publishing Group (NPG) and Europe PubMed Central (Europe PMC) provided data feeds. Repositories at the University of Edinburgh, Massachusetts Institute of Technology (MIT), Imperial College London, the University of St Andrews, and London City University were involved in testing transfers from the RJ Broker.

9. Proposed service documentation

Includes service offering, SLD and roadmap documents for the RJ Broker but also for ORI on which RJ Broker has a dependency.

10. Evaluation forms

Evaluation forms have been designed to capture user feedback from the data suppliers and data consumers of the RJ Broker, see Annexe 3: User Feedback Forms.

11. RepNet evaluation and sustainability plan

Participation and input in RepNet sustainability plan and evaluation framework, see Annexe 1: Logic Evaluation Planner and Annexe 2: UK RepositoryNet+ Generic Evaluation Question.

12. Knowledge built

Sharing know-how with RepNet partners through attendance and presentation at workshops and meetings.

13. Knowledge dissemination


3.2 How did you go about achieving your outputs / outcomes?

The objective of the RJ Broker project was to design, develop and test a standalone middleware tool to handle the delivery of research output from data suppliers to multiple repositories that would:

- Be a standalone middleware tool,
- Have both APIs and GUI,
- Accept deposit of research articles,
- Identify target repositories,
- Deposit the research article with registered repositories.

The project extension built on the initial objective to further the transition of the RJ Broker project to a potential service that EDINA or RepNet could run. The follow objectives were added:

- Setup of a service environment,
- Develop the functionality to support essential service to most IRs,
- Add selected new features identified as desirable during the first phase of the project,
- Target the registration of participants to the data trials to the future RJ Broker service.

All objectives have been achieved. However the registration of the participants to the data trials to the full RJ Broker service was unsuccessful. This sixteen month project followed the JISC project management guidelines and governance requested by the RepNet project. A project plan was drawn up at the start of the project which divided the work to be done into distinct work packages and set the milestones to be achieved. The activities carried out in each work package are reviewed below.

3.2.1 Project Management, Dissemination and Outreach

*Work Package 1(WP1) and extension 1 & 6 (WP1-ext & WP6-ext)*

**Project outputs:**

- Output 1: Project plans
- Output 2: Final report

Document title: EDINA RJ Broker Project Final Report
Last updated: December 2013 – v1.1


Reporting

The project plans identified four milestones to do with reporting: the production of the original project plan and extension work plan, a revised project plan, a risk & issue register and a final report. The project plan was submitted to Jisc and RepNet on the 30th April 2012 and a revised version improved to address the feedback received was produced by the end of May 2012. Funding for a four month extension was confirmed on the 7th May 2013, a work plan was submitted on the 17th May 2013 and with a final version a month later.

Project Board meetings that included the Jisc programme manager, a representative for RepNet project and the RJ Broker project manager were held monthly to review the progress of the project. The project manager circulated a Project Highlight Report to the attendees ahead of the meeting that summarised the work completed, status of the project, issues, risks and any changes to the project. The template provided by RepNet was used for these reports. No significant risk or issues were recorded during the life time of the project.

Additional progress reports and presentations were made on request, in particular for the RepNet Service and Innovation Partner Group (SIPG) and RepNet Board meetings.

Dissemination

RepNet aims to promote the UK repository infrastructure and as such dissemination included the work done by its component providers as one of their main activities. The RJ Broker benefitted from these efforts. The proximity of the RepNet project team within EDINA allowed the RJ broker team to effectively share knowledge and expertise throughout the development, testing and setting up of the infrastructure for RJ Broker.

Dissemination of the project work was also done by the RJ Broker team through the project page (2) and other EDINA dissemination channels such as the EDINA website (3), Newsline newsletter (4) and community report (5). The Open Access Repository Junction (OA-RJ) blog (6) was used to promote continuation of this previous project into the current RJ Broker project. Nine posts, mainly discussing aspects of technical developments, were published during the lifetime of the project and flagged to RepNet for linking to their blog.

The project team attended several international and national meetings including Open Repositories (OR2012) (7), Repository Fringe 2012 and 2013 (8), hack sessions and workshops run by Gateway2Research, Dev8D and the Open Knowledge Foundation, and several RSP events. Our active participation in these events resulted in direct contact with stakeholders and many fruitful discussions with interested parties both existing and new contacts.

In particular, we organised a standalone workshop during OR2012, taking advantage of many of the stakeholders to the RJ Broker service attending the main conference. The purpose was to discuss the proposed RJ Broker application and gather further functional requirements. The meeting was by invitation only to key stakeholders in order to promote valuable exchange of ideas. Twenty-two people attended and constructive feedback was received. The discussions were summarise in the blog post “RJ Broker: a Research Output Delivery Service” (http://oarepojunction.wordpress.com/2013/01/10/rj-broker-a-research-output-delivery-service/). Presentations at an RSP webinar on the 29th May 2013 (9) and event on the 12th June (10) were very effective at disseminating the project efforts, specifically the outcomes of the data transfer trial involving NPG and Europe PMC.

These efforts resulted in active interaction with potential data suppliers and with IRs interested in receiving content from the broker including:
Suppliers | Elsevier, eLife, Frontiers, NORA, Oxford University Press
---|---
IRs | The University of Leicester, the British Antarctic Survey, the University of Hull, the University of Edinburgh, the University of Glasgow, the University of St Andrews, Imperial College London, City University London, the University of Oxford, the University of Southampton

The RJ Broker live site and the documentation are available online at [http://broker.edina.ac.uk](http://broker.edina.ac.uk). As a result of these dissemination activities, awareness of the broker and its potential is high amongst the stakeholders. Efforts should be dedicated to capitalise on this and recruit data suppliers and consumers to use the future service.

**Evaluation**

The RJ Broker project participated fully with RepNet evaluation efforts. Two inputs were requested:

1. A logic evaluation planner form provided in June 2012 was filled in and returned in July 2012. It is included in *Annexe 1: Logic Evaluation Planner*

2. The answers to a further three generic evaluation questions were requested in June 2013 and are included in *Annexe 2: UK RepositoryNet+ Generic Evaluation Question*

In addition, forms to gather feedback from users and evaluate their experience of the RJ Broker have been designed and included in *Annexe 3: User Feedback Forms*. One form is aimed at data suppliers and the other aimed at data consumers. These can be made available on the RJ Broker website alongside the documentation for the APIs to capture volunteer feedback from users. Alternatively, the RJ Broker service provider can use them as part of their acceptance testing for the RJ Broker service following the addition of new suppliers or registration of new data consumers.

**3.2.2 Data Gathering**

*WP2*

**Project outputs:**

- Output 6: Template legal agreements
- Output 12: Knowledge built
- Output 13: Knowledge dissemination

Data from two external sources was used to test and validate the RJ Broker developments.

**Nature Publishing Group**

Participation of NPG (11) in the development and testing of the RJ Broker was initiated during the previous project on OA-RJ (12) and was sought following their involvement in the PEER Project (13). NPG is a major science publisher and a significant source of research output. NPG is willing to send records to the broker containing post print publications, including full text and all associated files such as data, diagrams or figures, at the time the manuscripts are being processed for publication. This would allow for a data feed to the RJ Broker to be integrated into the NPG workflow. However this also means that most of the records will still be under NPG standard six month embargo period. In order for RJ Broker to receive the records at publication time, NPG requested that a legal agreement be drawn to cover compliance with embargo periods.

Two legal agreements were necessary to cover the chain of responsibility:

1. Between NPG and EDINA to ensure that the broker respects any embargo period. This includes not exposing embargo data through the GUI and APIs, and only transferring embargoed records to IRs complying with the data consumer legal agreement.
2. Between EDINA and an individual repository to ensure that the IR respects any embargo associated with any record transferred by the broker.

The text of these agreements were written and approved through an iterative process between NPG and EDINA for the data supplier license, and between EDINA and MIT for the data consumer license.
MIT was selected as preferred partner in the trial by NPG. The agreement with NPG was signed by both parties on 23\textsuperscript{rd} August 2012, while the MIT agreement was in place on 10\textsuperscript{th} October 2012. Although these agreements were put in place to cover the testing period of RJ Broker, they can be used as template for future agreements that will need to be put in place in the future broker service. Further revisions may be required to cater for the specific needs of a given data supplier or IR. The templates have been shared with RepNet and various stakeholders following outreach activities. The feedback has been positive with IR managers willing to sign up to the future service. These agreements are available in \textit{Annexe 4: Legal Agreement Templates}.

\textbf{Europe PubMed Central}

RCUK and Jisc required the involvement of Europe PMC (14) in the data trial for the RJ Broker to focus the developments on the UK HE/FE stakeholders and support the UK funder mandates on Open Access (OA). Indeed the Wellcome Trust requests that any research output resulting from one of their grants should be deposited with Europe PMC. Late in 2012, the Europe PMC funders approved the developments required to push data made available via the Europe PMC web service directly to the broker as part of their 2013 Development Plan. Samples of Europe PMC data were provided to the RJ Broker. The access to this data is unrestricted and there was therefore no need for any legal agreement.

\textbf{3.2.3 Software Development, Essential Functionality & Additional Functionality}

\textit{WP3, WP3-ext, WP4-ext}

\textbf{Project outputs:}
- Output 3: RJ Broker middleware tool
- Output 5: Technical documentation
- Output 7: Metadata standard review
- Output 12: Knowledge built
- Output 13: Knowledge dissemination

\textbf{Development}

The development of the RJ Broker as a standalone middleware tool first required extraction of the broker functionality from the OA\_RJ, to establish it as a separate application and finally to extend its functionality to provide an end-to-end delivery facility. The proposed technical developments were for an EPrints based solution which would include:
- A repository to collect the records received from data suppliers,
- A set of EPrints importers to enable the deposition of records by data supplier to the RJ Broker internal repository,
- A set of EPrints exporters to enable the delivery of records to IRs by the RJ Broker,
- A set of bespoke scripts to manage the processing, routing, transfer and tracking of the records from data suppliers to IRs,
- A GUI & APIs to enable browsing and harvesting of the RJ Broker repository.

Although EPrints remains the most common repository platform in the UK to date, there was a requirement from stakeholders to include other repository platforms such as DSpace and Fedora during the project. Some of the necessary developments had started during the first phase of the project but more efforts were made to address these during the extension with the collaboration from IR managers involved in the use case data trials.
The functionality of the broker was initially developed to fulfill the requirements of the stakeholders gathered at the workshop in July 2012 and discussed in the blog post “RJ Broker: a Research Output Delivery Service” mentioned previously. The feedback from the data trial and input from the RepNet project were later also included. In particular we wish to mention work on:

1. **Metadata field selection**: The metadata provided by the RJ Broker is principally used to facilitate the transfer and easy upload of records. RJ Broker can include any field in the metadata it provides to IRs, for example to support a given standard, as long as the value for this field has been provided to the RJ Broker by the data supplier. The RJ Broker aims to support OA and its stakeholders and therefore the metadata supplied by the RJ Broker has been tailored to suit the needs of this community. These include funder and grant codes as advocated by the RIOXX guidelines, embargo details and ORCIDs. The RJ Broker metadata fields are detailed in the online documentation.

   In addition to the RJ Broker metadata, the full record, as provided by the data supplier, is transferred to repositories. This allows IRs to collect additional information from the record if this was given by the supplier but not included in the RJ Broker metadata. Including fields from the record to the RJ Broker metadata highlights the information that may be included in the record and also makes this information OA. Indeed, the metadata for a record is always OA while the record itself may be embargoed. However, it is worth noting that the RJ Broker is only a delivery service and whilst it will support standards, it cannot enforce their compliance.

2. **Coding and scope of embargo**: Preliminary work on the broker assumed that an embargo applied to the entire record. Discussion with stakeholders, and NPG specifically, highlighted that although some files may be embargoed, the metadata was not. The RJ Broker moved from a simple implementation of embargo with an internal dual archive store: one open and one “dark”, to a more granular model with a single archive store where metadata is always OA, but any individual file may be embargoed. Two technical posts have been written on this topic: “Embargoes in Real Metadata” (http://oarepojunction.wordpress.com/2012/12/11/embargos-in-real-metadata/) and “Embargoes in Real Metadata Take 2” (http://oarepojunction.wordpress.com/2013/02/14/embargoes-in-real-metadata-take-2/)

3. **GUI and APIs**: The ability to browse and download the data in the RJ Broker is extremely useful to encourage potential IRs to subscribe to the delivery service, as well as potentially interact with data consumers that may not yet or will never be able to support SWORD
delivery, such as Campus Research Information Systems (CRIS) or individual researchers. The EPrints GUI and OAI-PMH harvest facilities were customised to support search by institutions and repositories. A variety of metadata formats is supported as standard, the RJ Broker metadata profile has been added and the RIOXX could be too when its profile is released for EPrints. In addition an API for one off bulk transfer of data between the broker and an IR was implemented to enable newly registered IRs to receive the backlog of records that the RJ Broker holds for their institution.

4. **Two levels of service:** The development of the RJ Broker in the first phase of the project was for a delivery service to registered IRs with assumption that all registered IRs had agreed to respect embargo periods. The second phase recognised that legal agreements may become a hurdle to the take up of the potential service. Therefore the RJ Broker was modified to provide two levels of service:
   a. Delivery of all content available under OA to all repositories which have registered their SWORD details with the broker
   b. Delivery of embargoed files is added to OA metadata records where the IR has signed the legal agreement to honour embargo.

   This was achieved by developing a filter in the routing process where the legal status of the recipient is checked before an embargo record is forwarded to them. The OA level of service enables a quicker registration of IRs and will provide a staged approach to joining the service.

5. **Notification service:** In addition to the two levels of service, the ability to offer a third service to the wider IR community was implemented. This consists of a regular email notification service to non registered IRs where a summary of the records available for a specific IR is emailed to that IR manager with a link to the RJ Broker, as well as information on how to harvest the data and register for the delivery service. This development required modification to the ORI in order to expose the contact details harvested via the ORI GUI and APIs.

**Standard Review**

The standards used by RJ Broker, including metadata, data and transfer protocol standards, have been selected for their popularity and wide use to ensure an easy fit for the RJ Broker into existing research output data workflows to facilitate use and uptake. The RJ Broker also aims to promote the use of appropriate standards to support interoperability between repositories and compliance to funder mandates.

A review of relevant standards was carried out and a short report provided to RepNet and Jisc on the 18th January 2013, see Annexe 5: Review of Metadata Standards for Potential Support by the RJ Broker. The table below summarises our conclusions:

<table>
<thead>
<tr>
<th>Standard</th>
<th>RJ Broker support</th>
</tr>
</thead>
<tbody>
<tr>
<td>CERIF</td>
<td>RJ Broker supports the use of CERIF in the data. RJ Broker is currently designed to work with IRs rather than CRISs. It is likely that in the future interaction with CRISs will be desirable and that CERIF fields that are considered important can be flagged for addition to the RJ Broker metadata fields.</td>
</tr>
<tr>
<td>RIOXX</td>
<td>RJ Broker is RIOXX compatible and includes the fields identified by RIOXX in its metadata. RJ Broker intends to fully support RIOXX by installing the application profile to the RJ Broker EPrints application as and when the profile is released.</td>
</tr>
<tr>
<td>OpenAIRE</td>
<td>RJ Broker supports the OpenAIRE metadata fields in the data. The RJ Broker can support the OpenAIRE guidelines in the metadata if required.</td>
</tr>
<tr>
<td>SWORD 2.0</td>
<td>RJ Broker supports SWORD 1.3. A move to SWORD 2.0 is likely to be desirable in the future as IRs move to SWORD 2.0. However, due to the changes between the two versions, recent implementations of the standard often do not support backward compatibility to previous versions. This means that the broker would need to implement functionality to support both versions. This was documented in the blog post: <a href="http://oarepojunction.wordpress.com/2013/01/16/sword-1-3-vs-sword-2/">http://oarepojunction.wordpress.com/2013/01/16/sword-1-3-vs-sword-2/</a></td>
</tr>
</tbody>
</table>
Testing & Deployment

The middleware tool was to be delivered to RepNet for inclusion in their infrastructure and potential deployment as a service. RepNet were using the ITIL framework for the deployment of new services. We worked in close collaboration with RepNet to agree each step of the ITIL process. The RJ Broker was developed in a standard EDINA application container. The testing involved the creation of internal repositories (EPrints and DSpace) to act as targets for delivery. The migration process involved RepNet duplicating the installation first in a quality assurance (QA) container to undergo thorough testing and then to copy this installation to a live production site.

Figure 2: RJ Broker deployment path within the RepNet infrastructure

To assist with this, the application code and formal test suite were stored in the RepNet installation of the version control system github (15). Sample data records tailored for testing specific aspects of the broker were also included, i.e. publication for a given set of IRs. A technical installation guide was provided to allow RepNet to install, test and deploy the RJ Broker. This document underwent several iterations of feedback and improvements.

Documentation

The RJ Broker tool has been fully documented for use, support and development. Several technical documents have been produced to cater for these different needs and targeted at their respective audiences.

External documents available to all are:
- User Guide: http://broker.edina.ac.uk/information.html
- Data Supplier Engagement: http://broker.edina.ac.uk/supplier.html
- Data Consumer Engagement: http://broker.edina.ac.uk/repository.html
- Data Format Specification: http://broker.edina.ac.uk/data_format.html
- FAQs: http://broker.edina.ac.uk/faq.html

Internal documents are available to the development team and RepNet on Redmine – the web based application for issue management used by EDINA which includes software, issue and document tracking, and wiki facilities. For example, this includes the installation document mentioned in the previous section.

3.2.4 Data Trials, Use Cases, Scalability & Early Adopters

WP4, WP5-ext

Project outputs:
- Output 8: Use cases
- Output 12: Knowledge built
- Output 13: Knowledge dissemination

A set of data trials were performed with the multiple aims of testing the broker with actual real data, establishing connections with data suppliers and IRs, providing end-to-end use cases, assessing the performance and scalability of the RJ Broker installation, and recruiting users for the potential service.
**Data Suppliers**

NPG and Europe PMC agreed to the necessary developments to provide a data feed to the RJ Broker. Once participation in the trial was agreed, see section 3.2.2 Data Gathering, supplier engagement followed these steps:

1. **Agree a data format**: NLM DTD is the standard descriptive format used in the PEER project and adopted by many publishers. However, suppliers often use their own interpretation of the standard, or may use their own format which is often the case when the supplier is not a publisher. Therefore, it is necessary for a supplier to agree on the format for the records they will send to the broker.

2. **RJ Broker develops a bespoke importer script**: to accept and successfully process the incoming records from the supplier.

3. **Provide login credentials to the RJ Broker QA installation**.

4. **A set of test transfers** from the supplier to the broker are performed to troubleshoot any connection issues.

5. **Data feed test** to the RJ Broker test installation

Several months were needed for the RJ Broker and NPG to go through these steps. NPG had first to schedule the development works to integrate a data flow to the RJ Broker from their publication workflow. This was followed by a several tests and further technical development to ensure correct transfer of the records. Then in February 2013, the test data feed was established and 300 records were received. These contain full text, embargoed records, with NLM DTD based metadata. The trial was deemed successful.

The data feed between Europe PMC and the RJ Broker was set up quicker thanks to the participation of Mimas (16), another component provider to RepNet. Mimas proposed a 1 month work package to create a system which harvested data from the Europe PMC public API, and pushed it to the broker. Late in 2012, the Europe PMC funders approved these developments as part of their 2013 Development Plan. This work was scheduled for and completed in January 2013. This includes metadata only records with unrestricted access. The data feed was first tested in February 2013 and a transfer of 60,000 records over three days was done. This highlighted issues where multiple records are included in a single deposit. The performances of the broker and data profile of this trial were analysed in an internal report provided to RepNet and Jisc. A further data feed of 24,000 records was carried out in July 2013 to assess the improvement in processing bulk records. Both trials were successful.

These trials resulted in about 85,000 records gathered in the RJ Broker trial installation at [http://devel.edina.ac.uk:1203/](http://devel.edina.ac.uk:1203/). Using the affiliation provided within a record, RJ Broker was able to identify one or more organisations for 32,000 records. For these records with known organisation(s), RJ Broker identified one or more target repositories for 20,000 of these records. This content is spread across 904 organisations and 1,300 repositories worldwide.

**Data Consumers**

The first external repository involved in testing the broker was OpenDepot.org (17). OpenDepot.org runs EPrints 3.2 which is the most common repository software and version combination in the UK, and therefore a good representative case. It is also managed by EDINA which eased the development of the necessary data importer script. That script now serves as a template, to be used and customised for other EPrints 3.2 IRs. OpenDepot.org was useful in testing the broker with test records, but although it was included in the data trials, it is not an institutional repository and therefore was not a target for any of the 85,000 records.

MIT, as a preferred partner for NPG, was the second repository involved in the trial, and the first to enter the legal agreement. Their repository runs DSpace 1.8, an application that was not included in the original project scope. MIT were willing to help with the development of the data importer script needed to receive records from the RJ Broker into DSpace. The principal challenge was to include support for embargo in the metadata set accepted by DSpace needed to receive NPG records. This is
discussed further in section 3.3. Unfortunately MIT were not able to achieve this before the end of the project. Therefore, to date, no data has been transferred to MIT.

In parallel, RepNet through their engagement in the STARS shared initiative (18) worked on developing a DSpace importer for the repository of the University of St Andrews as part of their use case to demonstrate the use of the RepNet components in a research intensive institution. RepNet developed an importer script for DSpace 1.8 able to understand a subset of the RJ Broker metadata based on the standard DSpace import routine. In particular, this script does not include metadata fields about funding, as required by the RIOXX guidelines, or fields coding embargo details.

This script was used by the RJ Broker team to establish a successful connection between the broker and a test instance of the University of Edinburgh repository Edinburgh Research Archive (ERA) (19). Unfortunately, it was not suitable to set up a live connection to ERA because its content is driven by their PURE CRIS system and any content delivered directly to the repository would be erased by the daily overwrite from PURE.

However, this script can now serve as a template for use and customisation by other DSpace 1.8 IRs that do not wish to receive embargo records. More importantly, it can be used as a starting point for the development of a rich metadata importer script for DSpace as discussed in section 3.3.

The universities of Southampton and Glasgow were contacted and were keen to take part in the trial. Both have EPrints 3.3 repositories. EPrints 3.3 supports SWORD 2.0 while the RJ Broker supports SWORD 1.3. During the initial investigation in the development of a data importer script able to work with EPrints 3.3, it was found that the SWORD 2.0 implementation in EPrints 3.3 is not backward compatible and therefore no longer supports SWORD 1.3. This has been reported as an issue which is now scheduled to be fixed fix and included in an upcoming version of EPrints (see 3.3.3 for further details).

The University of Hull was also interested in receiving the RJ Broker content. They run a Hydra repository based on Fedora which is not SWORD enabled yet. They are considering the implementation of a SWORD endpoint in Hydra as part of their future developments. However this is beyond this project time frame. Information of how to harvest the data from the RJ Broker using OAI-PMH was provided.

Finally, Jisc and the Wellcome Trust (WT) supported the involvement of Europe PMC as a data supplier and were keen for this data to be included in an end-to-end use case involving multiple IRs. Their preference was for Imperial College and the University of Oxford. RepNet and the RJ Broker worked with Imperial College to establish a connection between their DSpace 1.7 IR and the broker. The data importer script developed for DSpace 1.8 was adapted to work with DSpace 1.7, and a successful connection was established. This exercise was very helpful in troubleshooting some of the issues with the mapping of metadata field between the broker and DSpace. However, Imperial College’s IR is a Full Text only IR, and like the universities of Edinburgh and St Andrews, relies on their CRIS to populate their IR. Therefore a live connection between the RJ Broker and their IR is not practical. The University of Oxford, which runs a Fedora repository, was contacted but similarly to the University of Hull, they have not implemented a SWORD endpoint in their IR. City University London was approached because their repository runs EPrints 3.2 for which a data importer script is available and they accepted to be part of the trial. A successful connection was established with their IR which was useful in providing the RJ Broker with feedback on metadata quality. Unfortunately, City University London also runs a Symplectic CRIS which drives the data input to their IR.

**Early adopters**

One of the aims of the RJ Broker trials was to recruit the participants into the future RJ Broker service. This was not achieved during this project partly because of technical issues as detailed above and partly because of the uncertainty surrounding the future of the service itself with the end of the RepNet and RJ Broker project.

However, many positive steps have been taken. NPG wishes to review the legal agreement that was drawn up for the trial period in order to adapt it to the new service. NPG also indicated that they were
in the process of adding ORCIDs in their metadata which could be available to the RJ Broker data feed in the future. This would improve the metadata and has the potential to help with the identification of target repositories. NPG also proposed to build the data feed incrementally, starting with two journals, and adding more journals as IRs sign up for the service. The journals could be selected to ensure the content is relevant to the registered IRs.

Europe PMC are willing to set up a live data feed to the RJ Broker service installation. They have also taken on board the feedback regarding the inclusion of the affiliation for all authors and the strong demand for full-text. Both of these require technical developments that need to be scheduled but could be available in the future.

None of the IRs where successful transfers were made were able to set up a live data feed from the RJ Broker because CRIS drive their IRs and/or had a Full Text policy.

**Scalability**

The data trials allowed the observation of performances and ability of the RJ Broker to cope with a large volume of real data. Monitoring scripts were put in place which showed that most records are dealt with within 2-3 seconds in the development environment. This indicates that the maximum number of records that could be processed per day is ~25,000. However, the load on the dedicated production environment will be less than that of the development one, and should provide better performance. On-going review of the performance and production setup should be done to ensure they match the expected throughput.

### 3.2.5 Service Environment Setup & Sustainability

*WP5, WP2-ext*

**Project outputs:**
- Output 4: Service infrastructure
- Output 5: Technical documentation
- Output 9: Proposed service documentation
- Output 12: Knowledge built
- Output 13: Knowledge dissemination

**Infrastructure**

RJ Broker fully engaged with the sustainability planning carried out by RepNet, as the expectation was for RepNet to operate the future service. However, following the Jisc review of the RepNet project, a new model emerged for the components to work independently under Jisc management. The implications of this were that the RJ Broker team were to set up and manage the production environment for the broker, and to manage the potential ORI service.

The broker team re-used the QA environment already set up by RepNet and set up the required production environment. EDINA is experienced in running live services and in meeting defined service levels. With that in mind a robust environment was put in place which includes backup, redundancy and load balancing between two geographically separate server rooms, with support infrastructure to keep them mirrored. This is discussed in the blog post "Setting up Redundancy on a Live Service" at [http://oarepojunction.wordpress.com/2013/07/03/setting-up-redundancy-in-a-live-service/](http://oarepojunction.wordpress.com/2013/07/03/setting-up-redundancy-in-a-live-service/). Technical documentation to install and manage the application is available internally.

The live RJ Broker is now available at [http://broker.edina.ac.uk/](http://broker.edina.ac.uk/) and ready to be moved into service.

**Service Documentation**

In order to support the transition to service efficiently, the following documents for the proposed RJ Broker and ORI services were written and shared with Jisc, RepNet and other component providers:

- Service Offering, see Annexe 6: Proposed Service Offering
3.3 Lessons learned

3.3.1 Fragmentation of IR Landscape
Although, EPrints and DSpace are common software for implementing repositories, there is a long tail of other applications available. All of these are available in several versions and all can be customised to the specific needs of an institution. This creates a very fragmented landscape into which to deliver a service.

It has proven difficult to obtain a clear vision of what this landscape is for the UK IRs. Information from sources like OpenAIRE, ORI or RepNet is not always complete or up to date. In an attempt to gather accurate information, direct machine-to-machine queries to all UK IRs listed in the ORI database were made in order to collect information about which software and version each IR used. The table below shows these results:

<table>
<thead>
<tr>
<th>Software (Total IRs)</th>
<th>Version</th>
<th>Number of IRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPrints (120)</td>
<td>2.3.*</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>3.0.*</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3.1.*</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3.2.*</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>3.3.*</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3.4.*</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>other &amp; undefined</td>
<td>75</td>
</tr>
<tr>
<td>DSpace (44)</td>
<td>1.6.*</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1.7.*</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1.8.*</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3.*</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>other &amp; undefined</td>
<td>25</td>
</tr>
<tr>
<td>Fedora (5)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>other &amp; undefined</td>
<td>4</td>
</tr>
<tr>
<td>Other (38)</td>
<td></td>
<td>38</td>
</tr>
</tbody>
</table>

Figure 3: Software & version in used in 207 UK IRs obtained from direct API queries on the 4th July 2013.

3.3.2 Support for Rich Metadata in DSpace
The developments leading to this RJ Broker project were mostly focussed on EPrints, with an early interaction with DSpace to confirm the basic concept was valid. The later, and more involved, work with DSpace has raised technical issues that are more complex than anticipated. For example, DSpace only assigned an identifier to objects when they are made available in the repository, not when deposits are received. The RJ Broker is therefore unable to track a deposit into a DSpace repository where those deposits go into a review queue, and consequently is unable to provide a link to the data supplier such that they can track deposit of this content.

The most crucial issue which was too big to be dealt with within this project is how support for rich metadata such as embargo can be added to DSpace. Indeed, DSpace has implemented two different models for the embargo process over time and these are mutually exclusive. The first implementation was coded by MIT and is available in older versions of DSpace like 1.8. It applies the embargo at the record level, both metadata and files. The newer implementation, coded by Atmire in newer versions (3.3 onward), is designed for their XML-based User Interface (UI), and provides more granularity with the ability to apply embargo at the individual file level.
The latest implementation is obviously more flexible and matches the data provided by the RJ Broker. However, this implementation includes a fundamental change to the underlying database which breaks the Java Server Pages (JSP) UI used in earlier DSpace versions.

Providing a solution in the JSP UI is not attractive because of the lack of granularity which will prevent us from making the metadata of embargo records open access. It is also a temporary solution which will become obsolete with upgrade to newer DSpace versions.

Providing a solution for the XML UI is better and will be future proof. However, it is not immediately relevant as to our knowledge no UK IR is using DSpace 3.3 yet.

A more complex solution which would apply to all DSpace versions would be to code a plugin that would allow the JSP UI to read the XML UI embargoes. This would provide in effect a compatibility plugin between the two DSpace implementations. We note that some work has started on this: https://github.com/DSpace/DSpace/pull/235.

### 3.3.3 EPrints SWORD 2.0 Implementation

The RJ Broker supports SWORD 1.3 and the newest version of EPrints support SWORD 2.0. Whilst SWORD 2.0 is able to replicate the behaviour of SWORD 1.3, the implementation of SWORD 2.0 in the current EPrints package (3.3.12) is not. This has been reported to the EPrints service as an outstanding issue in their tracking system which should be fixed in the next release (https://github.com/eprints/eprints/issues/106). When a solution becomes available, then RJ Broker will be able to interact with EPrints 3.3 repositories.

### 3.3.4 Full Text

The trial transfers and engagement with IRs have shown a significant shift towards full text only repositories. It has always been recognised that full text records were preferable but at the start of the project, IRs were more willing to accept any content, including metadata only records. Many IRs now only accept full text records. We see this as a positive move which will enrich the content of IRs and raise their values. The implication for the RJ Broker is that a new level of service is required to distinguish between full text records and metadata records, similar to the OA and embargo delivery levels. This has been included in the broker development roadmap. A further implication is on data suppliers to provide full text records if they want their data to appeal to IRs.

### 3.3.5 Changing Landscape

In the last few years, there has been a rise in the use of CRIS in UK institutions driven by the greater need for reporting. The main CRIS providers in the UK are Symplectic and PURE. The transfer trials have shown that a common setup for an institution is for the CRIS and IR to be closely integrated with content being accepted in the CRIS and later copied to the IR, usually via nightly overwrite. The delivery of content to their IR, which is what is offered by the RJ Broker, is not supported in this workflow. Therefore, there is a crucial need for the RJ Broker to deliver to systems other than IRs, and specifically to CRIS.

Early discussions with Symplectic and PURE indicate that neither support SWORD, the delivery protocol used by the RJ Broker. However, they would be willing to consider the implementation of a data feed from the broker via SWORD or another mechanism if there was a significant use case for it. They are also mostly interested in Full text records. Working with CRIS could be very effective for the RJ Broker, especially in tackling the complexity of working with many different IR applications and versions. However, this should be considered as an alternative offering as not all institutions have a CRIS. Compliance with an embargo by individual institution may also be harder to track when working via a CRIS rather than directly with an IR.
4 Conclusions
This project has successfully delivers what it aimed to achieved:

- the design and development of a standalone middleware tool that handles the delivery of research output between multiple data suppliers and IRs
- the demonstration of end-to-end use cases involving external data suppliers and IRs
- the setup of a production environment and preparation for transition to service

The result of these 16 months of technical developments is a service ready RJ Broker built on a stable framework to which further functionality can easily be added.

The effective dissemination efforts were well received and generated positive interest. They raised the awareness of the RJ Broker as a future service. This will ease the recruitment of users to the service in the next phase of work that will launch the RJ Broker service.

Although the engagements by both RJ Broker and RepNet have raised a number of technical issues, they also highlighted new and evolving practices, behaviour and expectations from stakeholders. This formed a much better understanding of the requirements and challenges ahead, not only for the RJ Broker service, but for other repository services too.

The next steps for the RJ Broker are:

- the service launch,
- the recruitment of data suppliers and consumers,
- the on-going support, maintenance and technical developments,
- A clear watch on evolving trends in the repository landscape.

5 References
6 Annexes

6.1 Annexe 1: Logic Evaluation Planner

6.2 Annexe 2: UK RepositoryNet+ Generic Evaluation Question

6.3 Annexe 3: User Feedback Forms
6.3.1 For Data Providers
6.3.2 For Institutional Repositories

6.4 Annexe 4: Legal Agreement Templates
6.4.1 For Data Providers
6.4.2 For Institutional Repositories

6.5 Annexe 5: Review of Metadata Standards for Potential Support by the RJ Broker

6.6 Annexe 6: Proposed Service Offering
6.6.1 ORI
6.6.2 RJ Broker

6.7 Annexe 7: Proposed Service Level Description
6.7.1 ORI
6.7.2 RJ Broker

6.8 Annexe 8: Proposed Service Roadmap
6.8.1 ORI
6.8.2 RJ Broker

6.9 Annexe 9: Helpdesk Workflow