ECP-2007-DILI-537003

PEER

D2.2 Final report on the provision of usage data and manuscript deposit procedures for publishers and repository managers

Deliverable number	*D-2.2*
Dissemination level	*Public*
Delivery date	*31 October 2009*
Status	*Final, 29 Oct 2009, v5*
Author(s)	*Barbara Bayer-Schur; Foudil Brétel; Natasa Bulatovic; Gabriella Harangi; Wolfram Horstmann; Friederike Kleinfercher; Rianne Koning; Vilius Kučiukas; Marianna Mühlhölzer; Dale Peters; Laurent Romary; Jochen Schirrwagen; Maurice Vanderfeesten*
Internal reviewer	*Christoph Bruch; Jacques Millet*
External reviewer	*CIBER group, UCL*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

1 OJ L 79, 24.3.2005, p. 1.

Table of Contents

Tables, Figures & Appendices

Tables

Figures

Appendices

Introduction

The *Draft report on the provision of usage data*[1] *and manuscript deposit procedures for publishers and repository managers*, deliverable 2.1, set out to establish a workflow for depositing stage-2 outputs in and harvesting log files from repositories to enable the research envisaged in the PEER project. As that report preceded the tendering process whereby the respective research teams were selected, a number of issues were flagged for attention, particularly of the Usage research team, in WP5 and have since been referred for consultation.

A significant outcome of the previous draft report was the recommendation to establish the PEER Depot as a closed intermediary repository, to receive publisher deposit in the form of both 50% of the full-text outputs, as well as 100% of the metadata outputs; and to serve as a base line control for the research process. The PEER Depot has since been established, and has come to play a significant role in the workflow developed. While the draft report set out a preliminary deposit workflow from publishers to repositories, the central role of the PEER Depot has since influenced further developments in the provision of usage data and manuscript deposit procedures for both publishers and authors.

This report is the result of an ongoing negotiation between stakeholder groups comprising publishers and the library/repository community to establish best practice in deposit procedures that are least disruptive of existing publication workflows, while minimizing additional effort in repository ingest activities.

1 Methodology

Interaction between stakeholder groups has been conducted in a series of face-to-face meetings, in which a progressively increasing number of participants from both publisher and repository communities chose to participate by teleconference. Not only does this signify more efficient communication, it also indicates a growing sense of trust amongst and between stakeholder groups, borne from a common understanding of project objectives, and a pragmatic understanding of the complexity of everyday work processes encountered by both parties.

The draft recommendations of D2.1 were tested in the course of these discussions. Queries that have arisen in areas of concern are indicated and some alterations to the workflow are formulated in this final report.

Following the establishment of the PEER Depot, a pilot phase for publisher deposit to the PEER Depot was conducted in M10+11, with satisfactory results. A pilot phase for deposit from the PEER Depot to the repositories and the upload of log files was conducted in M12. Theoretically, the trial workflow has now moved into production, pending the validation of individual publisher deposits following the resolution of specific problems encountered during the pilot phase.

Two major accomplishments of the combined effort of WP2/3 have been the establishment of a responsible embargo management procedure, now conducted centrally at the PEER Depot for both publisher and author deposit; and an author deposit workflow, developed in progressive scenario testing process.

Standardised workflow set out in this final report enables a core group of interoperable European repositories, capable in theory of accepting material deposited from third party publishers and authors, beyond the project duration.

---

1 The DoW originally names this task „Harvesting of log files“. Since the recommended practice was altered, it is preferred in this document to call it “provision of usage data”.

A further significant achievement of the joint effort of WP2/3 has been formalisation of the transfer from the PEER Depot to all partner repositories in a single simultaneous process, using the SWORD protocol. Not only is this a new application in the transfer of both metadata and full-text articles, it represents a limited percentage of unknown errors in the transfer process. The intention is to have all PEER content mirrored in all participating repositories, to achieve a critical mass, except where precluded for technical reasons. The application of the SWORD Protocol represents best effort at achieving maximum content.

2 Repository Task Force

The Repository Task Force has been successfully established with the following six participating repositories:

- PubMan, Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. (MPG)
  http://dev-pubman.mpdl.mpg.de/pubman/
- HAL, Institut National de Recherche en Informatique et en Automatique (INRIA)
  Centre pour la Communication Scientifique Directe (CCSD/CNRS)
  http://hal.archives-ouvertes.fr/
- Göttingen State and University Library (UGOE)
  http://repository.peerproject.eu:8080/jspui/
- BIPrints, Uni Bielefeld
  http://129.70.12.25/opus4/public/home
- Kaunas University of Technology, Lithuania
  http://peer.elaba.lt/fedora/search
- University Library of Debrecen, Hungary
  http://ganymedes.lib.unideb.hu:8080/udpeer/

In addition, a UK-based repository has been invited to join the task force to better reflect usage of predominantly English language content expected. Preliminary enquiries, however, indicate a reluctance to participate in the project, ostensibly on the basis of heavy workloads of repository managers, who furthermore do not benefit financially from the project.

3 Interaction between stakeholder groups

Partners and stakeholders across Europe hosted meetings of work package 2/3: STM, London (M2) & (M4); Elsevier, Amsterdam (M6); INRIA, Paris (M8); the SURF Foundation, Utrecht (M10) and the Max Planck Digital Library, Munich (M13).

This interaction has been supported by the constructive mediation of the Project Manager, who participates in WP2/3 listserv discussions as a representative of the publisher stakeholder group. Similarly, the interaction with the research teams is mediated by WP1, and the research manager is also included in WP2/3 listserv discussions. A recent further development of this interaction has been the establishment of a repository managers' listserv, to include the research manager and a representative of the Usage research team.

4 Relationship between work packages and dependencies

Concern was expressed in the draft report at the disjuncture of work schedules in related work packages, so that decisions taken on a technical level in WP2 regarding the specification of log files, for example, might later impact on the suitability of data provided to the Usage research team in WP 5. With the subsequent appointment of the CIBER group from University College London (UCL), selected by tender to conduct the usage research, it has become possible to communicate relevant issues via WP 1, Manage Research Process.

The benefit of the dependency acknowledged between WP 2/3 and WP 4/5 has been demonstrated in the recommendation of WP 5 to include a UK repository. Since much of the

content is in the English language, usage rates will be much improved by increasing the geographic coverage accordingly.

An attempt has been made to improve the mediated communication between related work packages firstly by means of a designated repository representative, and subsequently by means of a shared listserv of all relevant parties.

The relationship between work packages remains a high priority to ensure that identified dependencies are addressed and miscommunication remains limited. For example it is emerged after much uncertainty that no common mechanism devised for repositories in the preparation of usage log files can be applied to all publishers. Publishers are individually negotiating with CIBER regarding their log file provision, since they do not have a uniform set-up internally. Therefore, this report treats only publisher deposit to repositories, and the usage data subsequently gathered in repositories.

1 Content deposits from publishers to repositories

1.1 Convention

In the context of the PEER project, content refers to stage-2 manuscripts and is understood as peer-reviewed article manuscripts with corrections as accepted for publication, but prior to editing and formatting for publication.

A trial/pilot phase for publisher deposit to the PEER Depot was carried out in M10+11 the results of which were satisfactory. Participating repositories are now ready to receive stage-2 research outputs from publishers via the PEER Depot, following the expiration of an agreed embargo period.

1.2 Established workflows

In an ideal world, publishers could directly deposit their content to repositories. But considering the different technologies provided by repositories and the disparity of technologies implemented by publishers, it appeared that a centralised point of collection, known as the PEER Depot, would be best suited to gather content from publishers, before processing and final deposit to repositories and to KB's long-term preservation (LTP) depot on behalf of the publishers. The e-Depot at the Koninklijke Bibliotheek in The Netherlands was invited to act as a long-term preservation archive, without participation in the usage measurement. The e-Depot acts in similar role to the publishing industry, and is therefore well positioned to enable the development of workflow, guidelines and standards that will secure the long-term preservation of the project's content. The PEER Depot is hosted at INRIA with the responsibility for facilitating publisher deposit and dissemination to repositories and to the LTP Depot. The content is also retained in the PEER Depot, in case of processing or delivery errors. The PEER Depot receives 100% metadata and 50% full-texts of the publishers' content. Some publishers only participate in the author deposit aspect, thus providing only metadata. The metadata is held extant to provide a control mechanism for the comparative research processes of measuring the balance of the 50% deposit by means of author deposit. This depot shall not be another repository, but a dark archive (not accessible, nor searchable).

The PEER workflow (Figure 1) shows the expected parallel paths of publisher deposit and author deposit.

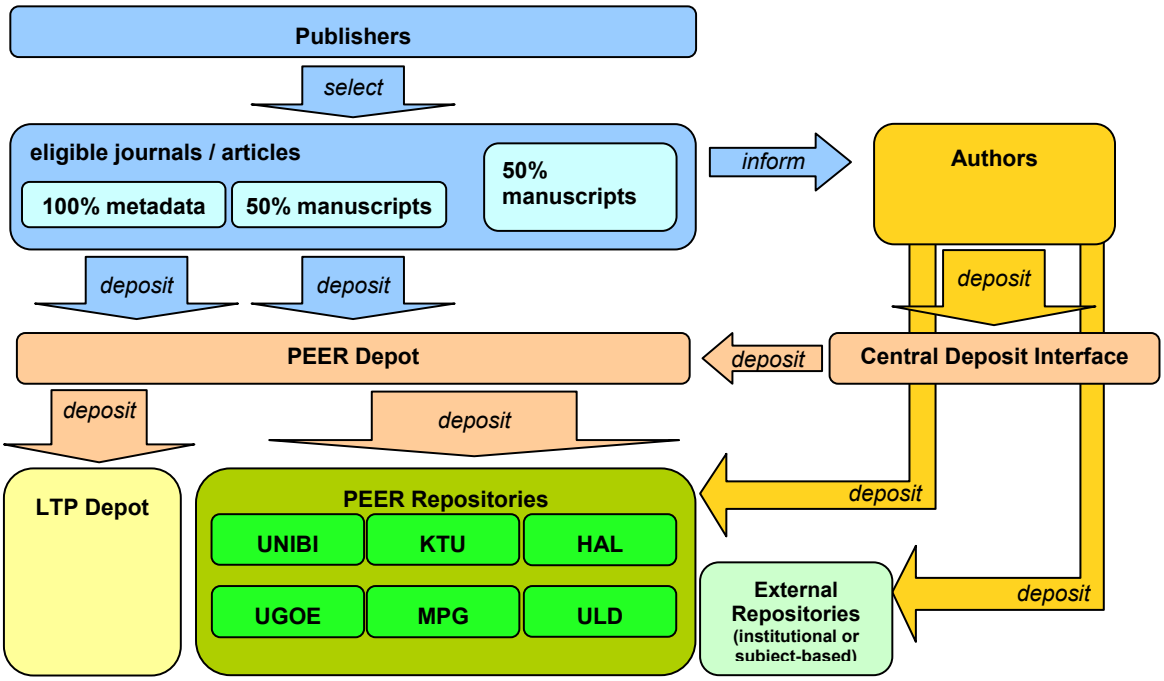

Figure 1: PEER workflow

1.3 Deposit procedures from publishers to the PEER Depot

Publishers deliver content (data + metadata) to the PEER Depot:

- On a daily basis or continuously
- Through FTPS or FTP into a dedicated directory
- As ZIP files, one per article
- File naming convention as [PublisherArticleId]_[yymmddhhmmss].zip [1]
- Preferably with an md5 checksum[2]
- The metadata file contained in the ZIP file should include the name of the full-text file, or the ZIP package must contain only one obvious full-text file.

Publishers provide in advance indication of:

- How to extract PEER-related metadata from the metadata file
- How/where to find the full-text in the zip file
- The deposit option chosen regarding metadata (see 1.3.2)

Publishers also provide in advance a list of journals contributed to PEER, with their assigned destination, i.e. publisher or author pathway (see Appendix A: *Participating journals*).

---

1 The PublisherArticleId may not be the same article-id as in the metadata, but it must be some kind of unique alphanumerical identifier. 'yymmddhhmmss' is the date in the form year in two digits, month, day, hour, minutes, seconds.

2 Each ZIP file should be delivered along with its checksum file.

1.3.1 Full-text format

For the sake of long-term preservation, the preferred file format of full-texts is PDF/A-1 [1]. Almost all publishers agreed to provide PDF (not PDF/A), which is also acceptable for the purposes of the PEER project. Publishers participating in the author deposit pathway do not provide the full-text in any format. Conversion of source files to PDF is not yet supported by the PEER Depot. The PDF file must include all figures. Provision of supplementary data is not needed since the PEER Depot does not forward them to repositories. Files indicating failed PDF conversion prior to transfer are excluded. The first provided version is authoritative over eventual following versions.

In order to identify articles, the full-text file received by the PEER Depot are renamed as follows: "PEER_stage2_[urlencoded-DOI].pdf" before submission by the PEER Depot to repositories and the LTP Depot.

1.3.2 Metadata

All publishers agreed to provide metadata in an XML format. Because every publisher uses a different DTD standard, it is decided that the PEER Depot would convert all publishers' XML into the TEI DTD standard. The TEI is a widely-used standard for encoding text materials in XML (including metadata). INRIA is in position to provide a 99,9% conversion transformation mechanism from any DTD to TEI.

Since exports to the PEER Depot might occur in different systems at different stages in the publication workflow, publishers indicated difficulties providing coherent stage-2 metadata. In some cases, critical metadata elements, such as embargo dates and persistent identifiers, are either added or first allocated at stage-3 in the publication workflow. To limit disruption of production workflows, it was agreed that the PEER Depot would support three options for gathering metadata. A submission is considered complete when all required metadata are provided.

- Option 1: All required metadata are submitted at stage-2 deposit.
- Option 2: Only a subset of metadata is provided during the first deposit including a publisher-article-id; the rest is provided in a second deposit during the embargo period including a publisher-article-id[1].
- Option 3: All the metadata updated by the publisher at stage-3 is submitted again, in *replacement* of the stage-2 deposit (except the document, which remains stage-2).

In option 2, for the second pass only, publishers can also provide the complementary metadata in the following forms:

a. a single XML file, not zipped
b. a CSV file (see Appendix B: *Technical specifications for CSV metadata provision*)

Derived from the DRIVER Guidelines [3], the minimum required set of metadata also includes the mandatory fields recommended in DRIVER viz.: Title, Creator, Date, Type and Identifier. Mandatory fields are marked (*).

While the PEER project recommends the submission of as much metadata as possible, the minimum requirements are marked (*) as set out below.

1 Because the second pass *completes* the first one, metadata provided twice are not updated.

DublinCore-like name	Comment
Title*	Article Title
Creator*	Corresponding Author's name: Last Name, First Name
AuthorEmail	Corresponding Author's e-mail address
Description	Abstract
Date*	Date of Publication
Identifier*	DOI or PublisherArticleId
Coverage	Geographic location of the Contributing Author: ISO 3166-1-A2
Journal	Journal Title
Affiliation	multi-tier organisation list: Country, Organisation, Laboratory
ISSN (e-ISSN, p-ISSN)	These elements are not mandatory to electronic publication and can be derived from CrossRef after DOI is provided. They may therefore not be provided by publishers.
Volume	
Issue	
First Page	
Last Page	
Type*	Default value = article. Mapped to info:eu-repo/semantics/article, info:eu-repo/semantics/acceptedVersion
Subject Headings	Subject headings; Scientific classification (defaults to what is provided in the PEER Journal tables)
Language	Language of the article, ISO 639-3 (defaults to 'eng')
Embargo	Embargo period for PEER Depot (defaults to what is provided in the PEER Journal tables)
Publisher name	Name of publisher (can be derived from the PEER Journal tables or FTPS homedir and is provided in the metadata file as an element)
Access	Open Access or Restricted

Table 1: Minimum metadata requirements

Since some articles may appear online only, or are published online before distribution of the paper edition, it was decided that the PEER Depot would not wait for missing metadata that should be provided by CrossRef (mainly *volume*, *issue*, *pages*), and transmit articles as

soon as possible. In this respect, the *volume*, *issue*, *pages* metadata can be considered as recommended, but not mandatory.

Finally, in the case of backfiles comprising previous articles, already set aside by publishers for the PEER project, and which might be delivered with only a DOI, but no further metadata, further investigation is required to source metadata from known public sources e.g. Public Library of Science (PLoS) or PubMedCentral. Each publisher will be approached individually to check whether backfiles can be provided in a format similar to current articles. In this case, ingestion to the PEER Depot and transfer to repositories can occur immediately, to facilitate the research process.

A database is used to store the metadata in the PEER Depot and to track events related to submission procedures (e.g. incoming and outgoing timestamps). This information can be made available to the PEER research teams, either through replication, or frequent exports. A complete list of articles processed in PEER is thus provided for comparative research between publisher deposit and author deposit procedures. The database also enables monitoring of the activity of the PEER Depot.

1.3.3 Embargo period

The period of embargo determines the date of distribution from the PEER Depot to participating repositories and to the LTP Depot. The duration of the embargo period differs from publisher to publisher and from journal to journal and also applies to author submission. These dates result in an agreed generic formula:

PublicationDate + EmbargoPeriod = Distribution Date

The publication date is provided in the minimum metadata set, defined either at stage-2 or stage-3 deposit. The embargo period, if not otherwise defined, defaults to that provided in the PEER Journal tables[1].

The embargo period on publisher contributed content is handled by the PEER Depot. For authors' content provided via the central deposit interface, the embargo period will also be handled by the PEER Depot (see Ch. 2.3.5). For publisher as well as author deposit the embargo period is applied according to the metadata provided in "date of publication" (see Table 1 above). As soon as the embargo period expires and the metadata file is complete, the content is ready to be transferred to and processed by the repositories and the LTP Depot.

1.3.4 Filtering

Two levels of filtering are envisaged as functions of the PEER Depot. Firstly, of journal titles by publishers for distribution to repositories and the LTP Depot, and secondly, of articles submitted by European authors. The PEER Depot receives 100% metadata and 50% publishers provided full-texts. All selected content − that is 50% of metadata and the corresponding full-texts − is disseminated to participating repositories and the LTP Depot.

The selection of publisher-deposited full-text is conducted at the journal title level, not manuscript level. The choice of eligible journal titles is defined by the publisher community, with due cognisance of research requirements, viz. behavioural response of specific subject disciplines.
See Appendix A: Participating journals

In addition, the filtering by type of non-research papers (i.e. letters to the editor) may be operated by the PEER Depot, if the Type metadata is provided.

---

1 See Appendix A & the project website http://www.peerproject.eu/about/participating-journals/

The project design further requires that only articles of European authors should be included in the study. Since publishers do not generally filter content in this manner, it was decided that the location of the corresponding author would be used to identify European content. The automated selection takes place at the PEER Depot, filtered against the coverage metadata element containing the geographical location of the corresponding author (by country). The contribution of additional European authors is regrettably lost to the research process.

An inevitable outcome of the project design, resulting from the filtering process is a limited research sample. While 50% full-texts of the publishers' content is disseminated to repositories and the LTP Depot, in fact, only that portion represented by the European corresponding author within that 50% are effectively disseminated. **The effective percentage of disseminated content will therefore be lower than 50%.** This issue is noted for further consideration, and possible adjustment of content quotas to ensure a valid research procedure.

1.4 Deposit procedures from the PEER Depot to repositories

A wide range of content formats submitted by publishers are normalised by the PEER Depot for transfer to participating repositories. Minimal metadata requirements for participating repositories are set out in the DRIVER Guidelines.[1]

- Participating repositories opt to set up a dedicated repository exclusively for receipt of PEER content; or to add content to an existing repository.
- Additional effort in the ingest of PEER content is limited to the implementation of the SWORD interface using the SWORD protocol (see Appendix C: *The SWORD protocol*).

The LTP Depot is not SWORD compliant, so for transferring the content from the PEER Depot to the LTP Depot, the FTP protocol or a FTP/s client will be used.

The deposit procedure uses a unified ingestion service, based on accepted international standards. These standards include PDF/A (ISO 19005-1:2005); TEI metadata format for descriptive metadata; ZIP for creating a package containing the PEER content; and the Atom Publishing Protocol (RFC 5023) using the SWORD specification as a transport protocol transferring the package to the repository. The benefit achieved is a core group of interoperable European repositories, capable in theory of accepting material deposited directly by third party publishers and authors beyond the project duration.

The deposit procedure is an automated process whereby the publications released from embargo are transferred from the PEER Depot to all partner repositories in a single simultaneous SWORD transfer. The intention is to have all PEER content mirrored in all participating repositories, to achieve a critical mass, except where precluded for technical reasons. When the publications in the repository are accepted and stored, an automated confirmation message is sent back to the PEER Depot with the online link to the publication. These locations can be used to notify the author about the links where he or she can find the stage-2 material.[2]

The deposit procedure from the PEER Depot to the repositories is illustrated in Figure 2 below.

1 DRIVER Guidelines v.2.0: http://www.driver-repository.eu/DRIVER-Guidelines.html
2 See minutes of PEER WP 2/3 meeting, 3rd September 2009, MPDL, Munich.

Figure 2: Deposit procedure from the PEER Depot to the repositories

1.4.1 Transfer procedures overview

The transfer of 50% full-text content and the author submitted files from the PEER Depot is conducted as follows:

- On a daily basis, as articles are normalised continuously
- Submission by FTP/S[1] transmission[2] or SWORD protocol
- As ZIP files, one per article[3]
- The ZIP package contains only one pdf data file and one metadata file
- File naming convention as
  - `[PEER_stage2_[urlencoded-DOI].pdf]`
  - `[PEER_stage2_[urlencoded-DOI].xml]`
  - `[PEER_stage2_[urlencoded-DOI].zip]`
    in order to identify PEER articles in repository log files, slashes in the DOI format are encoded as "_slsh_".
- Submission accompanied by an md5 checksum[4]
- In the case of FTP/S, an acknowledgement file named
  `ack_PEER_stage2_[urlencoded-DOI].txt`
  comprising only the repository internal identifier, which is the URL pointing to the created resource, will be returned in successful ingestion (void if unsuccessful).

1.4.1.1 Normalisation and Packaging

The metadata and the full-text files submitted by publisher deposit and that submitted by author deposit are normalised, since repositories expect a unified standard of the material. The different variations of the delivered metadata formats (mostly NLM format in different versions) are converted to the TEI metadata format. The full-text files are delivered by the publishers in PDF or PDF/A format. Both files are packaged in a ZIP compliant file.

The filename of these files are renamed to contain the DOI and follows the following syntax: "PEER_stage2_ [urlencoded-DOI].[ext]" for all files accounts where [ext] has to be replaced with respectively "xml" (TEI metadata), "pdf" (full-text) and "zip" (package). For all files accounts where [urlencoded-DOI] has to be replaced with the DOI string that accompanies the publication, and in some cases, not encouraged, all slashes in the DOI string may be replaced with the following string: "_slsh_" (for security reasons concerning the web server,

---

1 FTP/SSL is a secure way to transfer files. The opensource command line tool cURL can be used as a FTPS client.

2 FTP pull has two advantages: repositories do not have to install a FTP-server; and they have confirmation of successful ingest.

3 A single zip-file is essential to enable the PEER Depot to identify clearly each article, i.e. the material is not spread into many files that need to be gathered together.

4 Each ZIP file is delivered along with its checksum file.

when changing webserver configuration is not allowed). Then the filename is URL-encoded (RFC 3986), to avoid unusual behaviour upon unrecognised characters.

Examples using the DOI “10.2345/38884.299_299” creates the following filenames:

- PEER_stage2_10.2345_slsh_38884.299_299.xml
- PEER_stage2_10.2345_slsh_38884.299_299.pdf
- PEER_stage2_10.2345_slsh_38884.299_299.zip

The application profile of the TEI metadata tells the repository manager how to interpret the metadata fields in the PEER context. Both the TEI metadata DTD (see Table 1) and the way of packaging provide the repository a standard what to expect when they receive a PEER package. This standard is put under a unique namespace that can be used when sending the package using SWORD-APP. The name space goes by the URI: http://purl.org/net/sword-types/tei/peer/ .

Agreements and conventions

- Package contains one metadata file and one PDF file
- Package format is ZIP
- Metadata format is TEI (Text Encoding Initiative) according to the PEER-TEI Application Profile (see Ch. 1.3.2)
- PDF format is PDF/A (ISO 19005-1:2005)
- Filenames are renamed in the following syntax: PEER_stage2_[DOI].[ext]
- [DOI] is the Digital Object Identifier of the publication
- [ext] is the extension of the files, either PDF, XML or ZIP
- All the slashes in the filename may be replaced with: “_slsh_” This is not encouraged, the default action is to change the webserver configuration to allow slashes.
- All filenames are completely URL-encoded

1.4.1.2 SWORD: Transporting embargo released stage-2 material

The complete ZIP file is then ready for transfer to the repositories on expiration of the embargo period. The embargo period differs per journal and is listed accordingly in Appendix A: *Participating journals*. The algorithm setting the release date is described in Ch. 1.4.3 below. The transfer is authenticated via the SWORD-APP protocol, posts being authorised only by the PEER Depot.

Figure 3 below depicts the transmission action via the HTTP-protocol.

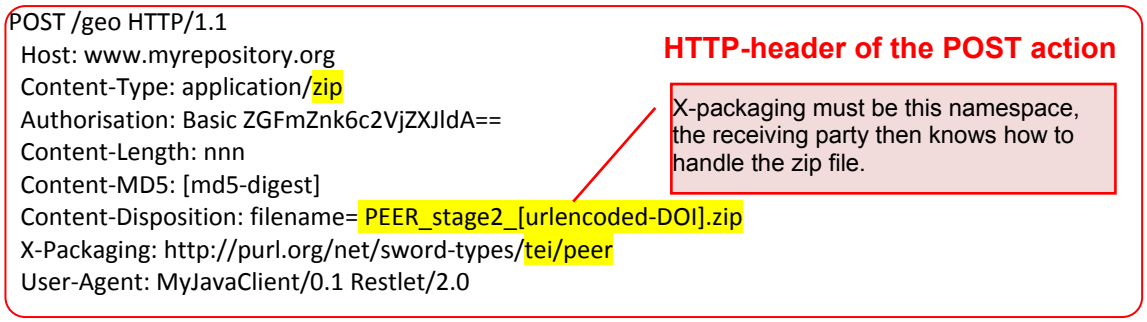

Figure 3: Transmission action via the HTTP-protocol

Agreements and conventions

- Submission accompanied by an md5 checksum[1]
- Basic authentication is used

1.4.1.3 SWORD: Notifying successful transfer with file location

When the SWORD interface has received the package it unpacks the ZIP-file and stores the PDF and Metadata into the repository. When this is done the SWORD interface immediately notifies the PEER Depot about the successful operation with the URL of the PDF located at the repository.

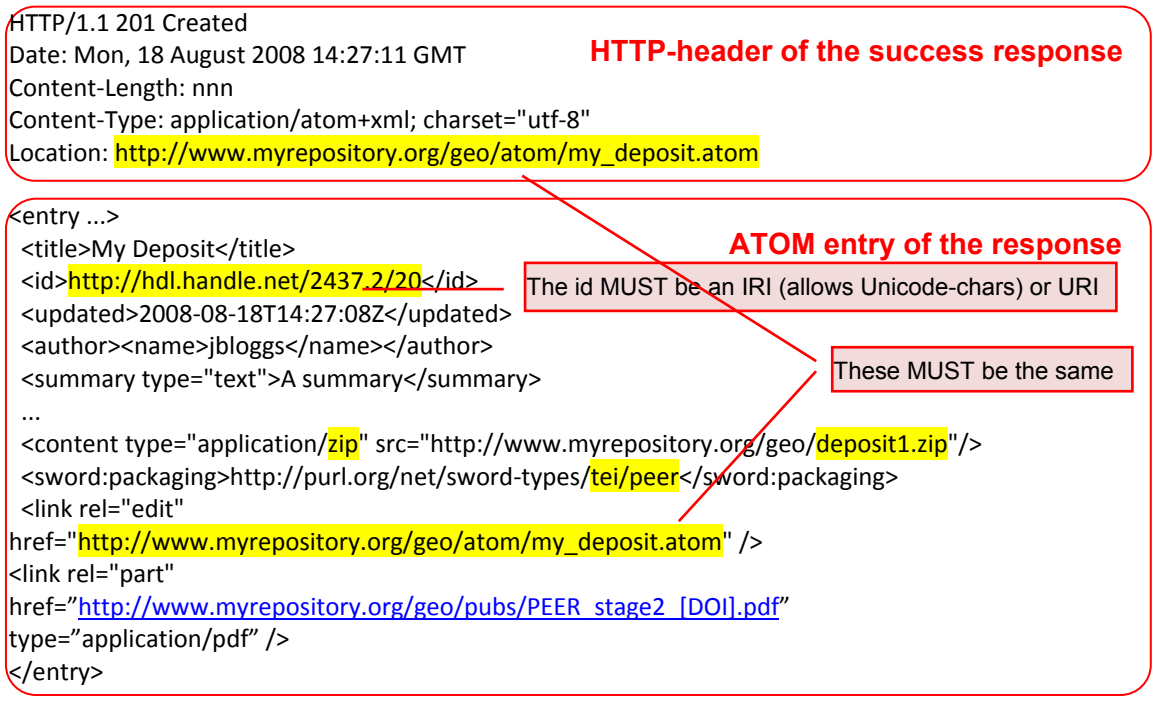

Figure 4: Notification of successful transfer from PEER Depot to repository

1 Each ZIP file is delivered along with its checksum file.

Agreements and conventions

- HTTP-header response element “Location” MUST contain the URI of the Media Link Entry, as defined in ATOMPUB.
- The Media Link Entry URI MUST dereference.
- The Media Link Entry URI MUST contain an <atom:content> element with a “src” attribute containing a URI.
- The Media Link Entry URI MUST contain the location of at least the PDF file in the repository.
- The Media Link Entry MAY occur more than once containing other relevant locations to the publication at the repository.
- The Media Link Entry URI MUST NOT contain internal server paths.
- <atom:id> MUST contain an IRI (Internationalised Resource Identifier, RFC 3987), allowing Unicode, or an URI (which is a subset of an IRI)
- <atom:author> MUST contain the user sending the package, it MUST NOT contain the author of the publication.
- Additional mandatory fields are <atom:title> and <atom:summary>

1.4.1.4 SWORD: Notifying unsuccessful transfer of the file

In the case of ingestion things might go wrong in three places:

1) At the HTTP protocol level
2) At the SWORD interface level
3) At the repository upon ingestion

Level 1 and 2 describe errors that happen on the surface, on the “communication” level. Level 3 describes an error that occurs below the surface, inside the repository.

Providing error handling at the HTTP level (1) is considered standardised in all repositories, this MUST be used, and will not be mentioned here further. Providing error handling at the SWORD interface level (2) described in the SWORD protocol v1.3 SHOULD be used, and will not be explained here, but we will refer to the SWORD v1.3 specifications. Error handling at the repository level (3) SHOULD also be used and will be explained below.

Ingestion Error feedback

When a file has been successfully transferred to the repository, the case might be that the repository cannot ingest the received file. The most appropriate response might be that there is something wrong with the ingestion and not with the transmission.

To provide the PEER Depot with a clue about that the file is not processed in the repository the following error handling information SHOULD be used.

Error URI	**Usage notes**
http://peerproject.eu/sword/error/ErrorOnIngest	The server MUST also return a HTTP status code, which describes the situation most likely.

This introduces a new namespace "http://peerproject.eu/sword/error" and an error type "ErrorOnIngest", which means the document couldn't be stored in the repository because of an error like

- the repository is down (status code 503)
- the repository takes too long to answer the request
- the repository requires authentication, the SWORD interface cannot fulfil

SWORD Error feedback

The following error handling procedures are written in the SWORD protocol v1.3, and SHOULD be implemented to provide useful error handling information.

Error URI	**Usage notes**
http://purl.org/net/sword/error/ErrorContent	The supplied format is not the same as that identified in the X-Packaging header and/or that supported by the server
http://purl.org/net/sword/error/ErrorChecksumMismatch	Checksum sent does not match the calculated checksum. The server MUST also return a status code of 412 Precondition Failed.
http://purl.org/net/sword/error/ErrorBadRequest	Some parameters sent with the POST were not understood. The server MUST also return a status code of 400 Bad Request.
http://purl.org/net/sword/error/TargetOwnerUnknown	Used in mediated deposit (see Part A Section 2) when the server does not know the identity of the X-On-Behalf-Of user.
http://purl.org/net/sword/error/MediationNotAllowed	Used where a client has attempted a mediated deposit, but this is not supported by the server. The server MUST also return a status code of 412 Precondition Failed.

Table 2: SWORD error feedback

Agreements and conventions

Due to rapid implementation, it has been decided[1] not to put effort in the error handling procedures. The first priority is to get SWORD working and rely on the standard HTTP error messages to identify problems.

However, it is recommended that the SWORD and Ingestion error handling is enabled to provide finer granular feedback. This information is useful for better analysis when a problem occurs, that might lead to a quicker solution.

1.4.2 Metadata

Publisher profiles indicate a wide range of metadata schema deployed. Derived from the DRIVER Guidelines[2], the minimum required set of metadata elements common to all publisher submissions, will be transferred to repositories:

- Mandatory elements : Title, Creator, Date, Type and Identifier
- Additional recommended elements as available
- PEER Depot transforms received metadata to TEI
- PEER Depot exports only TEI metadata files, as per repository preference

1.4.3 Embargo period

The embargo period differs according to each journal.

- Publication date plus embargo period determines the date of distribution from the PEER Depot to participating repositories

  PublicationDate + EmbargoPeriod = Distribution Date

  - Where “publication date” is the date of publication of the stage-3 publication, it can be found in the metadata provided by the publisher.
  - “Embargo period” is the period of time an article is not allowed to be released determined by the name of the Journal that can be found in the table of Journals participating the PEER project (see Appendix A: *Participating Journals*)
  - “Distribution date” is the date when the PEER Depot is allowed to transfer the article to the repositories (after the expiration of the embargo period).
- After author deposit, and if an e-mail address is provided, authors will receive a confirmation message that indicates notification of availability in the participating repositories following expiration of the embargo period. The confirmation message relies on the previous transfer of relevant metadata from publishers.
- Where possible, authors are then notified accordingly, with the links of the repository pages where they can find their deposited material.

1.5 Deposit procedures from the PEER Depot to LTP Depot

1.5.1 Introduction

The e-Depot of the National Library of The Netherlands (KB) aims to ensure perpetual access to the published records of the arts, humanities and social sciences, science, technology and medicine, and the digital cultural heritage. The KB assures publishers, libraries and end users

1 See minutes of PEER WP 2/3 meeting, 3rd September 2009, MPDL, Munich.
2 DRIVER Guidelines: http://www.driver-repository.eu/DRIVER-Guidelines.html

that the information preserved in the archive will outlast the transience of digital information carriers and formats. The role of the KB in the PEER project is to act as the long-term preservation (LTP) archive. The e-Depot is not an additional PEER repository, but in fulfilment of the curatorial responsibility of the library and repository community, will serve as long-term preservation depot in which the data objects and the accompanying metadata are kept safe beyond the duration of the project. The KB provides access to the content, based on the available access information in the metadata of the stage-2 manuscripts.

1.5.2 Content

The LTP Depot only receives and archives the final version of the content (PDF+ accompanying XML in one zip file) as delivered to the PEER Depot. The KB does not receive authors' content directly from the author, but (after a possible embargo period) via the PEER Depot with the authors' PDF incl. the accompanying complete (stage-3) XML metadata from the publishers. As the LTP Depot preserves content objects, only those records from the PEER Depot that contain an object and accompanying metadata will be transferred to and archived in the LTP Depot. This means that metadata only records are not transferred from the PEER Depot to the LTP Depot.

1.5.3 Workflow for Transfer to LTP Depot

The function of the LTP Depot is to preserve stage-2 manuscripts as deposited in the PEER Depot. Consequently, the LTP Depot has a different role and place within the PEER workflow functioning as an archive in which data objects and the accompanying metadata are kept safe. Figure 1 shows the place and role of the LTP Depot in relation to the PEER Depot and the PEER repositories.

As described in the workflow, stage-2 manuscripts are fetched from the PEER Depot by FTP/S. Before transfer takes place, the content of each zip file is converted to its final stage for archiving into the LTP Depot. Each zip file contains:

- A main file in full-text PDF format
- The complete accompanying metadata file

In processing the content, bibliographic metadata described according to the PEER/TEI DTD is converted to KB's DTD, whereas the original metadata delivered by the PEER Depot is stored with the content and converted metadata. Processing of the packages is based on the OAIS reference model[1].

The LTP Depot is investigating the possibility of responding an acknowledgement file named "ack_PEER_stage2_[urlencoded-DOI].txt" to the PEER Depot upon successful pre-process on its side. The acknowledgement file should comprise only the repository internal identifier (which may be a URL). Upon unsuccessful pre-processing, the acknowledgement file should be empty, and the rejection would be handled by the involved teams.

1.5.4 Metadata

Within the PEER project standardised metadata is applied according to the metadata requirements set out in Table 1. Publishers deliver the stage-2 manuscripts including metadata to the PEER Depot, and where possible, further recommended and optional elements are included. The PEER Depot converts the bibliographic metadata into the PEER/TEI DTD, which once transferred – as stage-3 i.e. most complete metadata − to the repositories and the LTP Depot is mapped according to local usage. The Dublin Core Metadata Element Set is commonly applied. Both the PEER Depot and the KB process

1 http://public.ccsds.org/publications/archive/650x0b1.pdf

bibliographic metadata according to the PEER /TEI DTD − stored in XML format − into their systems, together with the stage-2 manuscript. This bibliographic metadata may be converted to local workflows, for providing access through the local catalogue for example.

Further investigation is required of the possible future inclusion of an International Standard Name Identifier (ISNI) [4] and ultimately, a Digital Author Identification (DAI), [5] as these standards become more widely accepted.

1.5.5 Digital Preservation

PDF Guidelines

The article itself is required in a PDF-format. The KB maintains PDF guidelines which are mainly about the following subjects: Accessibility and structure, Fonts, Compression, Images, Executable actions and Colour. The KB is able to archive all PDF versions. For preservation purposes, PDF/A is the most suitable version. The main reason for this preference is that PDF files are portable across systems and platforms without changing the content or authenticity of the document now and in the future.

File information

For every file, main file and supplements, the KB requires the file name of a zip file, the file name of a main file, the file format and the file version. Regarding the file format the KB advises to handle PUID (Persistent Unique Identifier) as a standard. File information can be added in the metadata and is important for migrating files in the future, which might otherwise become unreadable. To ensure long-term preservation, it could happen that specific files need to be migrated to another (readable) file format.

The PDI (Preservation Description Information) is required for adequate preservation of the Content Information. Besides bibliographic metadata, the KB needs to identify metadata categories as specified under the OAIS model, listed below. For each category, the KB prefers a separate format. Currently the following categories and the related metadata format are preferred:

Category:	**Format:**
Bibliographic/Descriptive metadata	DCX
Structural Metadata	MPEG21-DIDL
Preservation Metadata	PREMIS
Provenance Metadata	
Technical Metadata	For still images: MIX For text documents: TextMD
Rights Metadata	For still images: MIX For text documents: TextMD

Table 3: Metadata categories specified under OAIS model

There are no strict boundaries between the different categories of metadata, some elements can also be sub-types of other elements and there is also a lot of overlap between the different categories.

References

[1] PDF/A-1 ISO 19005-1: Document Management − Electronic document file format for long-term preservation − Part 1: Use of PDF 1.4 (PDF/A-1). http://www.pdfa.org

[2] NLM Journal Publishing Tag Set http://dtd.nlm.nih.gov/publishing/

[3] DRIVER Guidelines
http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf

[4] International Standard Name Identifier: http://www.isni.org/

[5] DigitalAuthorIdentification:
http://www.surffoundation.nl/smartsite.dws?ch=eng&id=13480

2 Content deposits from authors to repositories

The development of an appropriate workflow for author deposits has proved to be most challenging, as the author response is unpredictable. This chapter sets out a process of author deposit that, as far as possible, does not interfere in established practice. Authors are therefore encouraged to follow their established practice of deposit in an institutional or subject-specific repository. Failing such practice, central deposit in the PEER Depot for distribution to designated repositories is recommended. It is highly unlikely that authors would be willing to deposit twice, nor does the project wish to impose additional work on those authors willing to participate. On the other hand, a direct author deposit procedure parallel to that of publisher deposit is not possible, without undue intervention in scholarly practice. Instead, authors eligible for participating in the PEER project are notified via the publisher and invited to respond.

There is currently no effective mechanism in place to ensure significant author participation, and without it, the value of the research might be questionable. The author deposit workflow is acknowledged as no more than an effort to keep track of authors self-archiving to PEER repositories and other repositories. However, it is precisely this lack of a controlled response to the author deposit procedures that will inform the behaviour and usage research investigations in Work Packages 4 & 5 respectively.

2.1 Options for authors

The author deposit procedure is envisaged in alignment with the normal points of contact between publishers and authors, as follows:

- Authors submitting manuscripts to eligible journals will be informed by the publisher about PEER and its objectives.
- At the point of acceptance, the author will be invited to participate, and to visit the PEER Helpdesk for further details of the project. The request for deposition will include a request to inform the project, should the author intend to deposit the manuscript in a repository of choice, other than in PEER (see Ch. 4.2.4.3).

2.2 Communication with authors

For reasons of data privacy, the participating publishers are not able to make the contact details of eligible authors available, and no direct communication is envisaged. Publishers are therefore provided with generic texts to communicate sufficient and consistent information to authors. At the point of acceptance of their manuscripts by their publishers, the authors will receive an invitation to deposit their manuscript within the framwork of the PEER project:

> This journal is participating in the PEER project <http://www.peerproject.eu/>, which aims to monitor the effects of systematic self-archiving (author deposit in repositories) over time. PEER is supported by the EC eContent*plus* programme <http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm>.
> As your manuscript has been accepted for publication by [Journal name], you may be eligible to participate in the PEER project. If you are based in the European Union, you are hereby invited to deposit your accepted manuscript in one of the participating PEER repositories. You may also choose to deposit in a non-PEER, institutional or subject repository in addition to, or as an alternative to deposit in a PEER designated repository. If depositing your accepted manuscript in a non-PEER repository, please set an embargo period of X months from the date of publication of the journal article for the public release of your accepted manuscript. For further information on PEER deposit, non-PEER deposit and embargo periods please visit the **PEER Helpdesk:** http://peer.mpdl.mpg.de/helpdesk.

However, since it is expected that authors may choose to respond immediately upon receipt of invitation to deposit, the invitation will be linked to the PEER Helpdesk website where authors are informed on their deposit options:

- For deposit to the PEER Depot, an online interface is established to guide authors through a simple deposit procedure (see Appendix D: *Peer Author Deposit interface specification*). In this case authors may provide their e-mail address for further contact by the PEER Depot upon successful deposit to participating repositories (see Figure 4).
- When depositing to other repositories (not participating in PEER), authors are invited to provide the URL of the item location in the repository, their name and optionally, an e-mail address for later contacts by the PEER research team. Although we cannot determine how many authors deposit outside the PEER Depot, because authors may or may not declare their intent, any information gathered on alternative deposit may provide useful to the behavioural research.

The PEER Helpdesk additionally provides further information to authors on the PEER project itself; information on participating repositories or on the handling of the embargo period. Authors may also post their questions via the *Trac*[1] ticketing system to the PEER project support team.
See Chapter 4: Ongoing support for publishers and repository managers

When depositing to the PEER Depot, the author receives two feedback messages:

1. Upon successful submission:
   A message is shown on the screen to notify the author that his/her submitted publication will be deposited in all PEER participating repositories[2] after the expiration of the embargo period.
2. Upon deposit to the participating repositories i.e. after the embargo period expires:
   The PEER Depot will transfer the author submitted manuscript to all participating repositories. As the SWORD[3] protocol is used for this purpose, each repository confirms the accepted deposit by a message containing the URL which indicates the location of the article in the repository. The URLs from all repositories are collected and e-mailed to the author.

2.3 Author deposit workflow

Several scenarios for the author deposit workflow were considered (see Appendix E: *Alternate author deposit workflow scenarios*) before the definition of the final workflow for the author deposit. As outlined in the [DoW], [D2.1] and [D3.1], it was assumed that authors would deposit their stage-2 manuscripts directly to the repositories participating in PEER.

A problem was foreseen however in the fact that this assumption does not take into account the authentication of the author with the repositories during the deposit workflow (see Ch. 2.3.1). As most participating repositories do not allow for anonymous deposits, some authors might not be able to deposit in the repository of their choice, even if they wished to do so. Some repositories allow for registration directly via their repository interfaces, but those repositories based at a university, are restricted by the authentication based on the network and the IP address of the client. Therefore, separate authentication of authors not affiliated to the repository host organisation would have been necessary for all repositories.

---

1 http://trac.edgewall.org/

2 Information on participating repositories is available both at the PEER project website (http://www.peerproject.eu/about/) and the PEER Helpdesk (http://peer.mpdl.mpg.de/helpdesk/wiki/repositorymanagers#PEERaffiliatedRepositories).

3 http://www.ukoln.ac.uk/repositories/digirep/index/SWORD

Due to data protection issues, the project is not allowed to use author e-mail addresses for authentication.

Possible alternatives to improve the author deposit − given that the anticipated low deposit rate threatens the validity of the project – are the registration at a central point and perhaps even a centralised deposit. The advantage of the latter is seen additionally in the possibility to enable any PEER author to deposit to each designated PEER repository, indirectly, through the PEER Depot. For this purpose, the SWORD protocol, originally applied to the transfer of publisher data from the PEER Depot to the participating repositories would also serve as the mechanism to facilitate author deposits to repositories.

2.3.1 Remote author authentication

The following alternative strategy for author authentication has been devised:

- Authentication at the PEER repository of choice, possibly by a single (PEER-guest) account. This account could then be used to disambiguate between the standard and the PEER related repository content.
- Centralised authentication conducted at the PEER Depot or the PEER Helpdesk, either by the author requesting an account (self-registration), by providing his/her e-mail address or as anonymous deposit (no authentication at all) – but with a spam-preventing functionality such as reCAPTCHA[1].

The recommendation for centralised authentication for remote author deposits found project support, though members of this work package are aware that this procedure requires extra effort to develop and integrate such functionality as a new application in the workflow.

2.3.2 Embargo management by repositories

The embargo period differs for each journal. A list of journal titles and the corresponding embargo period was provided by the participating publishers and is publicly available at the PEER website[2]. It has been acknowledged that repository management of the embargo period requires considerable and repeated effort. Therefore it was decided to manage the embargo of author deposits centrally at the PEER Depot prior to repository transfer in a manner similar to that for publisher deposit:

- Publication date extended for the duration of the embargo period determines the date of distribution of an article from the PEER Depot to participating repositories.
- PEER Depot holds any content previously received via author deposit until matching metadata are received from the publishers.
- Matching of publisher deposited metadata with the metadata received from author deposits determines the release of the deposits to participating repositories after the expiration of the embargo period.

2.3.3 Automated metadata matching process (duplicate author deposits)

To ensure the correct handling of the respective embargo period of an article, it was regarded necessary to conduct an automated process to match the author deposit with the corresponding metadata provided by the publisher at the repositories.

In the workflow originally envisioned, the PEER Depot would have transferred the metadata corresponding to an author submitted article only after the expiration of the embargo period signalling to the repositories the release of the article. The metadata provided by the author

1 http://recaptcha.net/
2 http://www.peerproject.eu/about/participating-journals/

would then have been overwritten with the publisher's version, since this is expected to be of a higher standard.

Repositories would have most likely received the full-texts from author self-archiving first, and thereafter the corresponding metadata from the PEER Depot (after expiration of embargo period). The identification would have taken place by matching author name and article title. Solely for the purpose of matching metadata and as an exception, taking the data protection issues into account, the use of author e-mail addresses was recommended as a means to match metadata and article. The DOI would not have been suitable as identification element, since the authors do not know the DOI at the time of deposit.

This procedure may still have resulted in some elements of manual checking e.g. author names written with special characters or variations of names (abbreviations, academic titles…).[1]

However, as both publisher deposits and author deposits are conducted via the PEER Depot, the process of matching of metadata has been natively moved to the PEER Depot. Authors provide the stage-2 manuscript metadata, full-text, and optionally corresponding author's e-mail, when making their deposit. This is the basis to match the author provided metadata with the metadata received from the publishers. Once metadata are matched and the embargo period has expired, the PEER Depot proceeds with the deposit of the stage-2 manuscript to the participating repositories. Thus an additional effort to match or overwrite author deposited metadata with the publisher deposits for repositories is avoided and the process is simplified.

2.3.4 Author deposit to a participating PEER Repository

In a process of consultation between members of the work packages 2 and 3, a series of author deposit workflow scenarios were developed.
See Appendix E: Alternate author deposit workflow scenarios

The guiding principle throughout remains the freedom of authors to choose to deposit their data to an alternative repository of their choice, in accordance with already established practice, and to inform PEER of this deposit. This functionality is enabled by the PEER Helpdesk.

Eligible (EU) authors who receive an invitation from the publisher to deposit their accepted manuscript to PEER are directed to the PEER Helpdesk, where they are offered two options:

1. Authors have the option to deposit their accepted manuscripts directly. Here they have the opportunity to enter their metadata and upload their manuscripts.
2. Authors may choose to deposit in their institutional repository, a subject-based repository or on their personal website. Authors who choose to do so are kindly requested to notify the project by inserting the URL of the article in the repository of choice (see Ch. 2.3.5 and 4.2.4.3).

There is no authentication mechanism in place; instead, a reCAPTCHA[2] is used to prevent automated deposits and spamming.
S*ee Appendix D: Peer Author Deposit interface specification*

The article and metadata submitted by the author are transferred to the PEER Depot where

a. the metadata is matched against those received by the respective publisher
b. embargo management takes place
c. the author is informed about transfer of data to repositories

---

2 See minutes of PEER WP 2/3 meeting, 25th June 2009, SURF, Utrecht.
2 http://recaptcha.net/

The chosen method of author deposit is regarded as a satisfying solution for both the project and the authors, since it limits the author's effort: By making one deposit the manuscript will be available in all participating repositories. The PEER author deposit workflow is described in Figure 5 below[1].

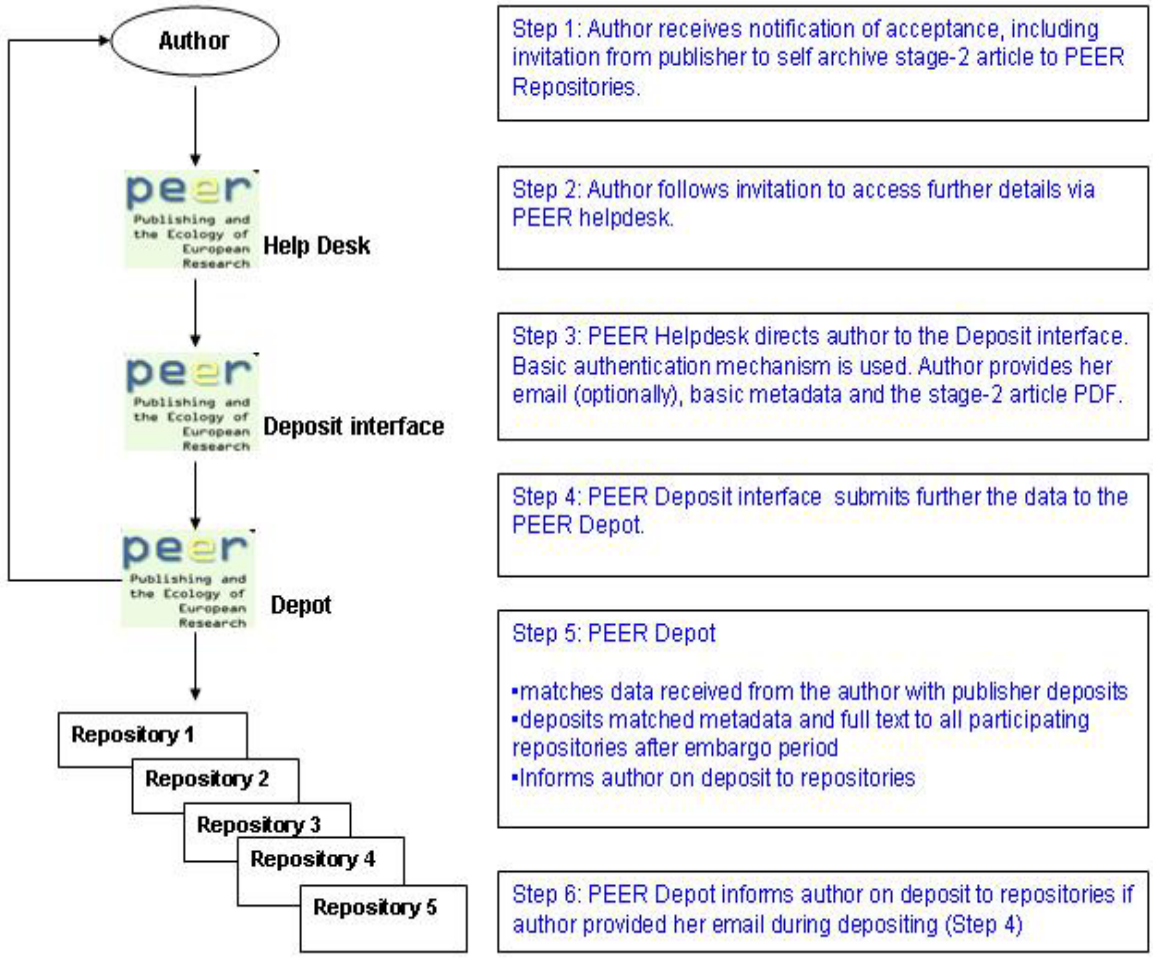

Figure 5: PEER author deposit workflow

2.3.5 Author deposit to a non-PEER repository

When invited to deposit data to PEER, authors are given an additional option to inform PEER on alternative deposit, in case they have deposited their stage-2 manuscripts in a non-PEER repository.

The PEER Helpdesk directs authors for this purpose to a form where they can provide information on the URL of the item/repository where they have deposited their data, their name and optionally, an e-mail address for further contact by PEER behavioural research team.

1 Further information is available at the Max Planck Digital Library Wiki (see http://colab.mpdl.mpg.de/mediawiki/Peer:_Author_Deposit).

2.3.6 Monitoring author response

Deposit will be monitored by the behavioural research undertaken in WP4, with the appointed team being the Department of Information Science and LISU at Loughborough University, UK, and measured against the 100% metadata control managed by the PEER Depot.

The project is aware of the fact that it is not possible to predict the behaviour of authors invited to deposit. It is noted that limited contact with authors, and hence minimal support for author deposit could affect the size of the research sample available in WP4. An option of supplementary harvesting by the PEER Depot, as a means of redress, requires further investigation.

3 Provision[1] of usage data

3.1 Introduction

This chapter defines how repositories should make available usage data to enable research on usage statistics, i.e. usage levels and patterns.[2] According to decisions taken in the PEER project, a very basic solution is presented: PEER repositories participating as usage data providers should upload usage log files in a regular manner. The usage log file is a text (ASCII) file containing, at a minimum, a record of the time and origin of requests for the PDFs provided by PEER.

The party selected to perform usage research in WP5 (CIBER group, see Ch. 3.1.2) is required to approach publishers individually for access to their log files. As a consequence, this interaction with publishers will not be described further in this report.
See Appendix F: Current and planned practice in the provision of usage data in a participating repository

3.1.1 Work package interdependency

The PEER project [1] will investigate the effects of the large-scale deposit of publications in repositories on user access, author visibility, and journal viability. Three tenders have been launched for behavioural and usage (December 2008) as well as for economic research (September 2009), respectively[3]. In order to enable this research organised in work package 1 and 5 of PEER, WP2 and WP3 are required to prepare the technical ground. This chapter describes basic assumptions and decisions relevant for specifying what WP2 and WP3 can provide for the usage research. The objectives of the usage research will be:

a. to determine usage trends at publishers and repositories
b. understand source and nature of use of deposited manuscripts in repositories
c. track trends, develop indicators, and explain patterns of usage for repositories and journals [2]

Thus, usage research requires:

- Complete information on the publications to be observed (see Ch. 2)
- Recorded usage events for these publications from all participating repositories as data-providers

The remainder of this chapter describes how these requirements can be met by the PEER project, specifically WP2 and WP3.

3.1.2 Usage research team

The CIBER group from University College London (UCL) has been selected to perform usage research. Together with the behavioural research team it will provide final reports

1 The DoW orginally names this task „Harvesting of log files“. Since the recommended practice was altered, it is preferred in this document to call it “provision of usage data”.
2 The publishers are individually reaching agreement with CIBER regarding their log file provision, since they do not have a uniform set-up internally.
3 http://www.peerproject.eu/press-releases-announcements/

mid 2011 and will feed into model development to determine whether (and how) traditional publishing systems can co-exist with self-archiving.[1]

CIBER is concerned that the rate of usage of the material be limited if no repositories from English speaking countries are included, since the vast majority of content is English language. They expressed the need to expand the repository task force to achieve better geographic representation. Thus, CIBER recommended the addition of a repository in the UK to better reflect the usage of predominantly English language content.

Furthermore, log files from the repositories for at least six months are required before PEER content becomes available in order to indicate if this additional content makes any difference to usage levels. For participating repositories that are dedicated PEER repositories this requirement cannot be met, since they contain no legacy content.

3.1.3 Motivation

Usage research in the domain of digital scholarly publications has recently been discussed intensively in the context of developing expressive indicators and metrics for the impact of scholarly publications (see [3] for a recent summary). Other than the conventional approaches based on citations and often related to complete journals rather than to the article level, usage events are thought to have the potential of providing higher temporal and thematic resolution ("quicker and more precise"). Methodologies have been developed [4], also in large scale projects (e.g. MESUR [5]) and standards are about to be expressed (e.g. PIRUS [3]). Within PEER, it was assumed that these developments are premature – thus implicating unforeseen work for the project – and it was decided [6] not to prepare the infrastructure for the use of such methodologies or standards but rather to provide 'raw' web-server log files to the party acquiring the usage research tender. Thus, specific questions to be answered by this document are limited as follows:

- How can raw web-server log files be transmitted from local data providers to the Usage research team?
- What is the structure of the log files?
- Which data shall be as minimum provided with the web-server log files?
- How can PEER articles be identified in log files?

3.2 Transmission of Log files

Local data providers upload their local log files to a secure server located at UCL. UCL has set up accounts for the data providers in order to upload by SSH based protocol rsync, SCP or SFTP. An automated upload by rsync over SSH on a daily basis with one (compressed) file per day is preferred. Alternatively the dropbox at <http://www.ucl.ac.uk/dropbox/> may be used. In this case the files should be sent weekly or monthly with all the daily files in a compressed archive format. The reader may picture this package as a tar.gz or zip-file with the naming convention:

PEER_usage_[data_provider_name]_[yyyymmddhhmmss].log.[tgz | zip]".

The chosen file naming convention is specifically designed to avoid mistaken file overwriting.

The KB, though, will not deliver log files to the PEER project.

1 Any data supplied to CIBER will be stored on a secure server located at UCL in London and held in accordance with UCL data protection policies, http://www.ucl.ac.uk/efd/recordsoffice/data-protection/

3.2.1 Structure of Log files

The log analysis team requests full and raw logs for two reasons [7].

1. Additional information over and above the minimum enables better validation of the data.
2. Additional information provides information on patterns of use, and thus the development of a richer model of user behaviour.

This implies refraining from the application of cleaning routines, typical of analytic tools such as AW-Stats [8]. Also, it is assumed (according to [6, 7]) that log files may contain non-PEER documents and that the filtering out PEER documents is an obligation of the research team (see also “Identification of Documents”) [11].

A generic and basic specification of log file formats is provided by the W3C [9], commonly used as “Common Log file Format [10]” and elaborated as “NCSA combined” or “NSCA extended”.

attribute	mandatory/ optional	example	Comment
host	m	125.125.125.125	maybe anonymised, see below[1]
rfc931	o	-	
username[2]	o	jdoe	
date:time	m	10/Oct/1999:21:15:05 +0500	Local time
request	m	"GET /PEER_stage2_10.1017_S1751731109003917.pdf HTTP/1.0"	PEER filename a must
statuscode	m	200	
Bytes	o	1043	
referer	m	http://www.google.com/	Highly recommended
user_agent	m	"Mozilla/5.0“	

Table 4: Log file format

Optional fields that are missing must be represented as “-“. Log files are ascii-textfiles. Fields are blank-separated and events are paragraph-separated. Please refer to the Website [11] for details.

An example is:

```
66.249.66.5 - - [12/Jan/2009:20:31:53 +0100] "GET /pdf_frontpage.php?source_opus
=87&startfile=Egelhaaf_et_al_UniForsch2002.pdf HTTP/1.1" 302 414 "-" "Mozilla/5.
0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
```

1 It should be noted that, at least according to German law, IP-addresses are not allowed to be recorded and handed over to a third party. IP-logging can be either suppressed in the configuration of the applied logging-routine or the log file has to be made anonymous before submitting them.

2 PubMan repository reports it cannot provide a username in the logs. This is also anonymised, and only logged-in users vs. non-logged in users are tracked.

As an apache configuration:

```
"%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\""
```

Where for reasons of confidentiality data has to be suppressed or anonymised then the redacted fields should be replaced with a hash value. In the particular case of IP addresses it is essential to provide the first three octets of the IP, e.g. '128.40.47.21' may be rewritten as '128.40.47.xxx'. A hash of the full value in addition is highly desirable.

The suggested procedure for redaction is thus:

Original log entry:

```
128.40.47.21 - - [31/Jul/2009:19:01:13 +0200] "GET
/docs/00/27/02/65/PDF/maladiesdesfemmes.pdf HTTP/1.1" 200 13032
"http://www.google.com/m?q=que%20se%20passe%20t%27il%20lorsqu%27une%20personne%20porte%20un%2
0gono%20qui%20n%27est%20pas%20bien%20soign%c3%a9&client=ms-opera-mini&channel=new"
"Opera/9.60 (J2ME/MIDP; Opera Mini/4.2.14881/504; U; fr) Presto/2.2.0"
```

applying a hash function on the IP address (The type of hash, e.g. MD5 or SHA, is not important)

```
hash_function('128.40.47.21') -> '1b5e84c1f858d5b9b6b06e47b6ca35ec'
```

Log entry provided for UCL:

```
128.40.47.xxx - - [31/Jul/2009:19:01:13 +0200] "GET
/docs/00/27/02/65/PDF/maladiesdesfemmes.pdf HTTP/1.1" 200 13032
http://www.google.com/m?q=que%20se%20passe%20t%27il%20lorsqu%27une%20personne%20porte%20un%20
gono%20qui%20n%27est%20pas%20bien%20soign%c3%a9&client=ms-opera-mini&channel=new"  Opera/9.60
(J2ME/MIDP; Opera Mini/4.2.14881/504; U; fr) Presto/2.2.0"
IPHASH=1b5e84c1f858d5b9b6b06e47b6ca35ec
```

It is expected that different software environments (e.g. simple apache server logs as in the case of standard repository systems or complex service oriented architectures as in the case of the MPDL) will cause different local policies for providing log files and some pitfalls are manifest:

- The filename or identifier appearing as http-request in the log file may only be known to the application (repository) but has no reference to PEER documents.
- In other cases raw log files may contain only cryptic calls of services (e.g. a PHP script[1] Web-services, Session Management, Cookie etc.) that does not contain any identifier and render a later identification of documents difficult or impossible.
- When http-'post' is used instead of http-'get' the identifier may be used as 'TYPE=HIDDEN' and does not appear in the log file.

These cases – as well as the many others that can occur – would render it impossible for the research team to infer which usage event belongs to a specific PEER-document. Thus, specific elaborations of the log files are to be prepared by an individual data provider. These elaborations might have different formats and encodings (e.g. TXT, CSV, XML, XLS) but the use of simple ASCII-textfiles is highly recommended, to avoid errors in the post-processing by the research team and to limit their workload.

3.3 Identification of documents

Raw log files contain much data of no relevance to the PEER project. Although WP1 have decided to leave the task of filtering out that data that are relevant for PEER to the Usage research team, it is the assumed responsibility of WP2 to indicate the identification of the usage

1 192.168.47.11 - - [15/Jan/2009:07:35:06 +0100] "GET /sendfile.php?type=0&file_id=8c49d37b913076c63054db5414d545c0 HTTP/1.1" 200 61846.

events for publications relevant for PEER. This is conceived here essentially *as any kind of object identifier that can be used to match strings in the usage log files*.

As agreed [7], the research design (WP1) foresees that 100% metadata for publications eligible in PEER are provided by the publishers (via a continuous FTP upload to the PEER Depot). These metadata will be provided by PEER to the Usage research team (see 3.1.2), in order to obtain the current list at any given point in time, enabling the matching between usage events in log files and eligible articles.

It has also been agreed [12], that an identifier will be created at the PEER Depot (see Ch. 1.3.1) that should be used by repositories as a filename after the document has been received from the PEER Depot.

This *filename* of the full-text provided by PEER should, in an optimal situation, allow easy tracking of usage events in the log files. It is therefore mandatory for participating repositories to represent this PEER-filename, either in the URL of the document or in any form that allows a later mapping between an internal identifier occurring in the usage event and the PEER-filename[1]. In the latter case, the deposit procedures of a participating repository must thus ensure storage of the PEER-filename as an additional identifier for each document. Furthermore, the participating repository must provide a list with pairs of local identifiers and PEER-filenames or a pattern to match to the research team, to track which usage event belongs to which document.

It was also decided [6;7] that only 50% of the articles eligible in PEER are deposited on behalf of the publishers while the other 50% are subject to spontaneous author submission. The latter 50% will be deposited via a central author deposit interface to the PEER Depot and transferred to all participating repositories. Therefore the PEER Depot makes this matching before depositing to repositories. Thus usage events can be readily identified in the raw log files also for spontaneous deposited articles.

3.4 Expected Result

The expected result of this procedure is a service provided by UCL, by which each participating repository uploads raw server log files that contain usage events of PEER articles. The additional requirement of a list of articles eligible in PEER is subject to the specification of the deposit process.

References

[1] PEER – Description of Work.

[2] Calls for Research Tender, 22 December 2008
http://www.peerproject.eu/press-releases-announcements/

[3] http://ie-repository.jisc.ac.uk/250/1/Usage_Statistics_Review_Final_report.pdf

[4] http://arxiv.org/PS_cache/cs/pdf/0605/0605113v1.pdf

[5] http://www.mesur.org

[6] PEER Steering Committee Meeting, Frankfurt 28-Nov-2008.

[7] PEER Kick-Off Meeting, Sophia-Antipolis 12-Sep-2008.

---

1 The filename will be “PEER_stage2_[url-encoded-DOI].zip” for publisher content and “PEER_author_[url-encoded-DOI].zip” for author content. There may be an additional mapping between the original PEER filename and the filename used in the repository.

[8] http://www.awstats.org

[9] http://www.w3.org/TR/WD-logfile

[10] http://www.w3.org/Daemon/User/Config/Logging.html

[11] http://publib.boulder.ibm.com/tividd/td/ITWSA/ITWSA_info45/en_US/HTML/guide/c-logs.html

[12] PEER Technical Meeting, London 7-Nov-2008.

4 Ongoing support for publishers and repository managers

4.1 Introduction

Communication between the publisher community, the PEER Depot and the repository community has been ongoing during the course of the project and is documented in this chapter to bring together the recent developments and resolution of outstanding issues described in D2.1 *Draft report on the provision of usage data and manuscript deposit procedures for publishers and repository managers*.

Due to the overlapping nature of the work, the main point of contact between the members of work package 2/3 is the list service <peer-wp2-3@inria.fr>. The Project Manager serves to represent the publisher community and is included in the listserv communication mechanism. Face-to-face meetings are held regularly between participating publishers, the repository task force and above all, the members of the respective work packages. This provides the opportunity to discuss issues in detail which would exceed the limits of e-mail contact. Several meetings dedicated to discussing technical issues in work package 2/3 were held at various locations. Partners and stakeholders across Europe hosted these meetings: STM, London (M2) & (M4); Elsevier, Amsterdam (M6); INRIA, Paris (M8); SURF Foundation, Utrecht (M10) and Max Planck Digital Library, Munich (M13).

The draft recommendations of D2.1 were tested in the course of these discussions. Queries that have arisen in areas of concern are indicated and some alterations to the workflow are formulated in this final report.

4.2 Establishment of a Helpdesk

4.2.1 Helpdesk functions

Actors involved in the ongoing support facility envisaged include authors, publishers, repository representatives and PEER researchers. Support for stakeholders on deposit is available from two sources. Firstly, the PEER website offers general information on the project and detailed information tailored to the needs of the various stakeholder groups.

Secondly, a PEER Helpdesk[1] online interface has been established. The Helpdesk is envisaged as a key point of contact for all the stakeholder communities participating in PEER and has been established primarily as a central point of author support, available at <http://peer.mpdl.mpg.de/helpdesk>.

This online interface, linked from the PEER website, is an authoritative source of information. Publishers refer authors to the Helpdesk that, in turn, will direct the author to the deposit interface (see Ch. 2.3.5 & Appendix D: *Peer Author Deposit interface specification*). As an online interface the Helpdesk will facilitate outreach and information provision activities and will moreover provide advice and support on the implementation of the D3.1 *Guidelines*[2], and questions of deposit and transfer, as described in this report.

The Helpdesk will offer direct support by means of an online query and mediated response service throughout the project duration. Automated systems have been investigated based on the following project criteria:

- Meeting the diverse needs of three identified stakeholder communities
- Efficient query handling and response mechanisms
- Handling of specific query behaviour on predetermined information-seeking tasks

1 http://peer.mpdl.mpg.de/helpdesk
2 http://www.peerproject.eu/reports/

The technical representatives of work package 2/3 came to the result to implement the support facility in the form of a ticket system (software *Trac*, which offers a Wiki and an issue tracker in one). A ticketing system is highly effective since the questions and answers are well documented. Each query result will be published, and the participants are able to review arising issues. This system then also provides a mechanism of passive interaction for those seeking assistance, but are unwilling to ask – a notable online query behaviour pattern. Where the ticketing system is made public, the "wisdom of crowds" principle can be applied to gain more efficient response to complex problems. Furthermore, frequently asked questions (FAQs) have been developed on the basis of the query results for future reference and published on the Helpdesk site, based on that established in DRIVER[1]. Figure 6 below shows the generic ticketing system workflow:

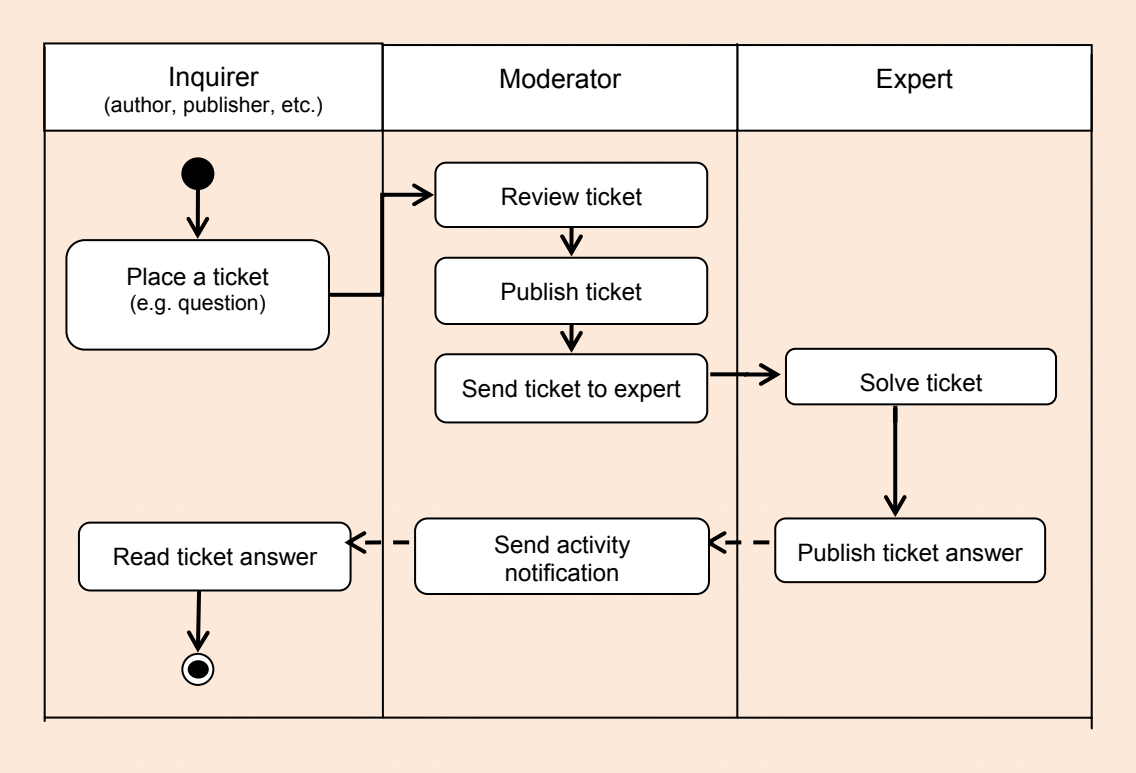

Figure 6: UML activity diagram of Helpdesk ticketing system workflow

4.2.2 Helpdesk Workflow

The Helpdesk workflow has been modelled on the DRIVER Helpdesk system, supported by the collaborative Wiki concept: Everybody receives all the queries and answers. But to ensure that each question gets answered, every query is passed on to a designated member of work package 2/3 who will be responsible for allocated areas of support. The moderator at SUB Göttingen is to monitor the Helpdesk and refer queries to representatives of WP2/3 as per designated responsibilities. Figure 7 below shows the Helpdesk input flow.

1 DRIVER Helpdesk: http://helpdesk.driver.research-infrastructures.eu/

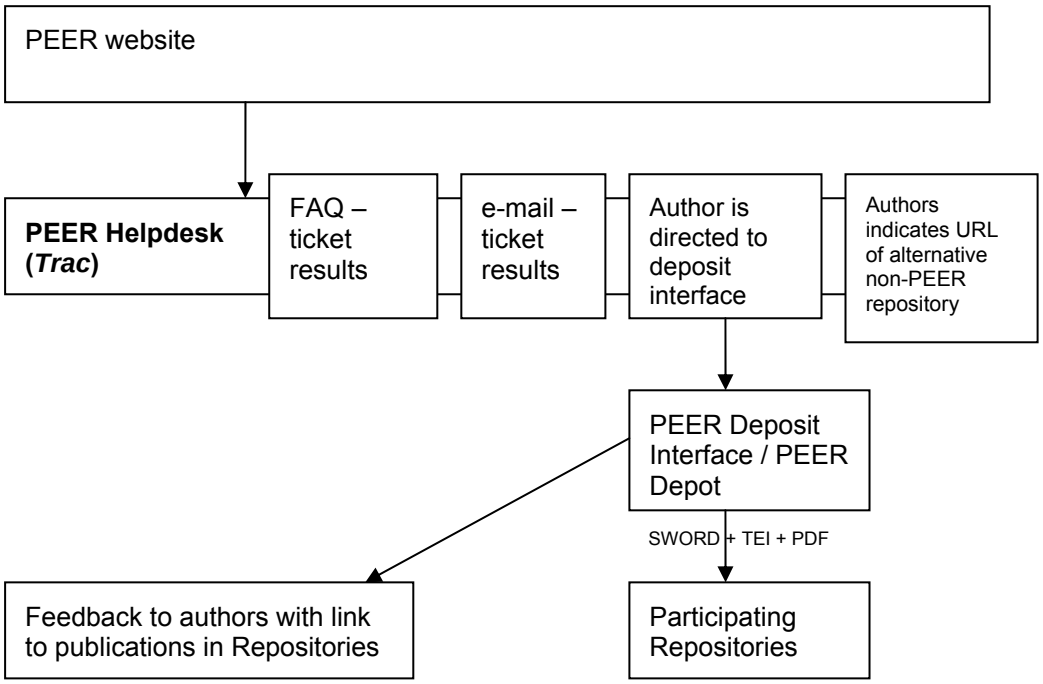

Figure 7: PEER Helpdesk: Input flow

4.2.3 Helpdesk for Publishers and Repository Managers

Publishers will deposit both 50% of the full-text outputs, as well as 100% of the metadata outputs from eligible journals at the PEER Depot. The 50% full-text outputs will be transferred from the PEER Depot to the repositories participating in PEER.

Although it is expected that implementing the D3.1 *Guidelines* is straightforward, the PEER Helpdesk will, however, support the consistent explanation and information on guiding publishers and repository managers through the deposition process.

The support for publishers is provided by experts resp. representatives of INRIA and is expected to cover queries regarding the expected metadata schema, transfer procedures, deviations from profile submitted, etc.

The support for repository managers is provided by expert representatives from SURF and MPDL. It is expected to cover queries regarding how to obtain the “NSCA combined” log file format, if not directly available. This might entail the provision of scripts for mapping from other formats, help for using the PEER-filename in the repository, advice on the corresponding interface to implement the SWORD protocol, etc.

4.2.4 Helpdesk for Authors

4.2.4.3 Guidance for authors on deposit procedures

For reasons of data privacy, no direct communication is envisaged between the project members and eligible authors. The procedures of author communication in the framework of the PEER project are described in detail in the D3.1 *Guidelines*. Therefore, the D3.1 *Guidelines* are not directed at the author community directly, but rather they reflect the considered opinion of the work package in consultation with the publisher community on recommended practice in offering assistance to authors.

The PEER Helpdesk will not only offer guidance to publishers and repository managers, often already involved in large-scale archiving, but mainly to authors, who may need

guidance in self-archiving for the first time. A recent study by Swan indicates that a substantial proportion of the author population (36%) are unaware of the possibility of providing Open Access to their work by self-archiving, and that only 49% of the author population have self-archived in some way. Of relevance to the PEER Helpdesk is the observation that authors have frequently expressed reluctance to self-archive because of the perceived time required and possible technical difficulties in carrying out this activity. However, similar findings suggest that only 20% of authors found some degree of difficulty with the first act of depositing an article in a repository, and that this dropped to 9% for subsequent deposits.[1] Therefore, information on the Helpdesk is intended to be given in a plain and easy way.

The authors will be guided to the PEER Helpdesk by their publishers. The PEER Helpdesk offers:

- Generic information on the PEER project
- Option of deposit
- The possibility to pose questions by creating a ticket
- FAQ
- Information on embargo periods
- Information on and for publishers
- Information on and for repositories

Enabling the author to choose to deposit their data within the scope of the project but also to an alternative repository of their choice, in accordance with already established practice, the PEER Helpdesk offers the author two options of deposit:

a. PEER Author deposit

In line with the way of author deposit the project agreed upon (Ch. 2.3.4) the author will be guided from the Helpdesk to the central deposit interface for deposit of accepted manuscripts from participating PEER journals in PEER repositories. At the same time he/she will be alerted that by deposting once, his/her manuscript will be available in all the participating repositories after the expiration of the embargo period.

b. non-PEER Author deposit

As agreed in D3.1 *Guidelines*[2], the project will address those authors who do not wish to deposit within the framework of the PEER project. Accordingly, the author will be guided to a dedicated page on the Helpdesk, where he/she is requested to insert the URL of the article submitted in his/her preferred repository of choice. The data gathered will be sent to the behavioural research team.

No direct communication between the project and the author is envisaged for reasons of data privacy[3]. Nevertheless, for the project, in particular for the behavioural research team, it would be worth collecting details of non-PEER deposits. Therefore, the author is additionally requested to provide his/her name and e-mail address to enable the behavioural research team to possibly contact him/her with a questionnaire.

1 SWAN, A. & BROWN, S. (2005) *Open access self-archiving: An author study.* http://cogprints.org/4385/

2 D3.1 *Guidelines*, Appendix C, Ch. 2.1, http://www.peerproject.eu/reports/

3 D3.1 *Guidelines*, Ch. 4.3.1, http://www.peerproject.eu/reports/

The PEER Privacy Policy states:

> *If author information is to be used […], permission will be asked of the authors at the point where the author provides this information.*
>
> *Individuals may be contacted for the purposes of PEER research either by publishers, repositories, or directly by the research teams. Where such contact takes place, it will be undertaken in accordance with the data protection and privacy policies of the relevant organisation.*[1]

Hence, it is optional for the author to provide the two last details on a voluntary basis. The author will be alerted that by giving his/her e-mail address he/she agrees to be addressed by the behavioural research team. Although we cannot determine how many author deposits will be done outside the scope of the project, because authors may or may not declare their intent, the information gathered on alternative deposit may provide useful information to the behavioural research.

1 PEER Privacy Policy: http://www.peerproject.eu/privacy-policy/

5 Conclusions

This report concludes the development of an overall framework for depositing stage-two outputs in and for harvesting log files from repositories. An innovative workflow has been devised to describe and standardise the deposit from publishers to repositories that demonstrates, in a core group of interoperable European repositories, the capability of accepting material deposited from third party publishers and authors beyond the project duration.

The development of an appropriate workflow for author deposits has proved challenging, as the author response is unpredictable, and cannot readily be standardised. The guiding principle adopted is that authors are encouraged to follow their established practice of deposit in an institutional or subject-specific repository. Failing such practice, a central deposit in the PEER Depot for distribution to designated PEER repositories is recommended.

A number of concerns remain, and may yet impact on the project outcomes. The author deposit workflow described in this report is acknowledged as no more than an effort to keep track of authors self-archiving to PEER repositories and other repositories. While it is precisely this lack of a controlled response to the invitation to participate that will inform the behaviour and usage research investigations in Work Packages 4 & 5 respectively, without significant author participation, the value of the research may ultimately be compromised.

Another concern arising from the project design is the limitation of the research sample, resulting from the filtering process. While 50% full-texts of the publishers' content is disseminated to repositories and the LTP Depot, in fact, only that portion represented by the European corresponding author within that 50% are effectively disseminated. The effective percentage of disseminated content will therefore be lower than 50%. This potential deficiency is noted for ongoing monitoring in work package 3, and adjustment of content quotas is recommended during the course of the project to ensure a valid research procedure.

Evident too is the delayed research implementation affected by the 6 month embargo period. Ongoing attempts to secure back files accumulated by publishers may serve to alleviate this concern. Furthermore, log files from the repositories for at least the previous six months are required before PEER content becomes available in order to indicate whether additional PEER content makes any difference to usage levels. For participating repositories that are dedicated PEER repositories this requirement is invalid, since they contain no legacy content.

Despite the accepted recommendation by CIBER to support an increased rate of usage of predominantly English language content material by the inclusion of repositories from English-speaking countries, this has yet to be achieved. Although this recommendation will be pursued, preliminary enquiries indicate a reluctance to participate in the project, ostensibly on the basis of heavy workloads of repository managers, who furthermore do not benefit financially from the project.

The final report on the provision of usage data and manuscript deposit procedures for publishers and repository managers reflects a collaborative effort between publishers and the library and repository stakeholder communities to achieve a feasible workflow for depositing stage-2 outputs and for harvesting log files from repositories. The limitations of the project design have been identified and made known to the behavioural and usage research teams in WP 4 and 5 respectively, to monitor their anticipated impact on the project outcomes.

Appendix A. Participating journals

PEER: Author submission Journal list by publisher

Publisher/ Journal	ISSN	Broad Classification	Embargo* (months)	Language (if not Eng)
BMJ Publishing Group				
Journal of Neurology, Neurosurgery and Psychiatry (including Practical Neurology)	0022-3050	Medicine	6	
Journal of Medical Genetics	0022-2593	Medicine	5	
Sexually Transmitted Infections	1368-4973	Medicine	5	
Cambridge University Press				
The Journal of Agricultural Science	0021-8596	Life Sciences	12	
Bilingualism: Language and Cognition	1366-7289	Social Sciences & Humanities	12	
Journal of Biosocial Science	0021-9320	Life Sciences	12	
Journal of Helminthology	0022-149X	Life Sciences	12	
Science in Context	0269-8897	Social Sciences & Humanities	12	
Urban History	0963-9268	Social Sciences & Humanities	12	
Elsevier				
Annales d'Endocrinologie	0003-4266	Life Sciences	18	French
Annales de Dermatologie et de Venereologie	0151-9638	Medicine	18	French
Annals of Pure and Applied Logic	0168-0072	Physical Sciences	18	
Applied Acoustics	0003-682X	Physical Sciences	24	
Biomass and Bioenergy	0961-9534	Physical Sciences	24	
Blood Cells Molecules and Diseases	1079-9796	Medicine	18	
Brain and Language	0093-934X	Life Sciences	18	
Cell Calcium	0143-4160	Life Sciences	12	
Computers and Geotechnics	0266-352X	Physical Sciences	24	
Energy	0360-5442	Physical Sciences	18	
Enfermedades infecciosas y Microbiologia Clinica	0213-005X	Medicine	18	Spanish
European Journal of Radiology	0720-048X	Medicine	18	
European Journal of Soil Biology	1164-5563	Life Sciences	18	
European Journal of Surgical Oncology (EJSO)	0748-7983	Medicine	12	
Fire Safety Journal	0379-7112	Physical Sciences	24	
Immunology Letters	0165-2478	Life Sciences	12	
International Journal of Antimicrobial Agents	0924-8579	Medicine	18	

Journal of Pragmatics	0378-2166	Social Sciences & Humanities	24	
Journal of Theoretical Biology	0022-5193	Life Sciences	18	
Materials Science in Semiconductor Processing	1369-8001	Physical Sciences	24	
Nuclear Engineering and Design	0029-5493	Physical Sciences	24	
Radiotherapy and Oncology	0167-8140	Medicine	18	
Sociologie du Travail	0038-0296	Social Sciences & Humanities	24	French
Solar Energy	0038-092X	Physical Sciences	24	
Telecommunications Policy	0308-5961	Physical Sciences	18	
IOP Publishing				
Classical and Quantum Gravity	0264-9381	Physical Sciences	12	
Journal of Physics A: Mathematical and Theoretical	1751-8113	Physical Sciences	24	
Journal of Physics: Condensed Matter	0953-8984	Physical Sciences	12	
Nature Publishing Group				
Bone Marrow Transplantation	0268-3369	Medicine	6	
Embo Journal, The	0261-4189	Life Sciences	6	
Gene Therapy	0969-7128	Life Sciences	6	
Genes & Immunity	1466-4879	Life Sciences	6	
Leukemia	0887-6924	Medicine	6	
Nature Genetics	1061-4036	Life Sciences	6	
Nature Structural & Molecular Biology	1545-9993	Life Sciences	6	
Oncogene	0950-9232	Life Sciences	6	
Oxford University Press				
Family Practice	0263-2136	Medicine	12	
Molecular Biology and Evolution	0737-4038	Life Sciences	12	
Systematic Biology	1063-5157	Life Sciences	12	
Annals of Occupational Hygiene	0003-4878	Medicine	12	
Sage Publications				
Active Learning in Higher Education	1469-7874	Social Sciences & Humanities	6	
Concurrent Engineering	1063-293X	Physical Sciences	12	
Cultural Geographies	1474-4740	Social Sciences & Humanities	12	
Ethnicities	1468-7968	Social Sciences & Humanities	24	
European Journal of Cultural Studies	1367-5494	Social Sciences & Humanities	18	
European Journal of Industrial Relations	0959-6801	Social Sciences & Humanities	12	

European Journal of Women's Studies	1350-5068	Social Sciences & Humanities	18	
European Union Politics	1465-1165	Social Sciences & Humanities	24	
Global Social Policy	1468-0181	Social Sciences & Humanities	6	
Group Processes and Intergroup Relations	1368-4302	Social Sciences & Humanities	18	
Health	1363-4593	Social Sciences & Humanities	12	
History of Psychiatry	0957-154X	Social Sciences & Humanities	12	
International Journal of Damage Mechanics	1056-7895	Physical Sciences	12	
Journal of Biomaterials Applications	0885-3282	Physical Sciences	12	
Journal of Plastic Film and Sheeting	8756-0879	Physical Sciences	12	
Journal of Thermoplastic Composite Materials	0892-7057	Physical Sciences	12	
Public Understanding of Science	0963-6625	Social Sciences & Humanities	18	
Second Language Research	0267-6583	Social Sciences & Humanities	24	
Time & Society	0961-463X	Social Sciences & Humanities	12	
Vascular Medicine	1358-863X	Medicine	12	
Springer				
Biotechnology Letters	0141-5492	Life Sciences	12	
Cancer Chemotherapy and Pharmacology	0344-5704	Medicine	12	
Celestial Mechanics and Dynamical Astronomy	0923-2958	Physical Sciences	6	
European Journal of Clinical Microbiology & Infectious Diseases	0934-9723	Life Sciences	12	
European Journal of Epidemiology	0393-2990	Medicine	12	
Holz Als Roh und Werkstoff	0018-3768	Physical Sciences	12	German
Journal of Ornithology	0021-8375	Life Sciences	12	
Journal of Molecular Modeling	1610-2940	Physical Sciences	12	
Neophilologus	0028-2677	Social Sciences & Humanities	6	
Nonlinear Dynamics	0924-090X	Physical Sciences	12	
Queueing Systems	0257-0130	Social Sciences & Humanities	24	
Rheumatology International	0172-8172	Medicine	12	
Taylor & Francis Group				
Applied Economics Letters	1350-4851	Social Sciences & Humanities	18	
British Journal of Guidance and Counselling	0306-9885	Social Sciences & Humanities	12	

Civil Engineering and Environmental Systems	1028-6608	Physical Sciences	12	
Communications in Statistics – Theory and Methods	0361-0926	Physical Sciences	12	
Ergonomics	0014-0139	Physical Sciences	12	
International Journal of Environmental Analytical Chemistry	0306-7319	Physical Sciences	12	
International Journal of Psychology	0020-7594	Life Sciences	12	
International Journal of Remote Sensing	0143-1161	Physical Sciences	12	
International Journal of Systems Science	0020-7721	Physical Sciences	12	
Journal of Engineering Design	0954-4828	Physical Sciences	12	
Journal of Modern Optics	0950-0340	Physical Sciences	12	
Journal of Natural History	0022-2933	Life Sciences	12	
Journal of Sports Sciences	0264-0414	Social Sciences & Humanities	18	
Optimization Methods and Software	1055-6788	Physical Sciences	12	
Phase Transitions	0141-1594	Physical Sciences	12	
Philosophical Magazine Letters	0950-0839	Physical Sciences	12	
Phsychotherapy Research	1050-3307	Social Sciences & Humanities	12	
Wiley-Blackwell				
Applied Cognitive Psychology	0888-4080	Social Sciences & Humanities	24	
Applied Organometallic Chemistry	0268-2605	Physical Sciences	24	
Biomedical Chromatography	0269-3879	Physical Sciences	24	
Biopharmaceutics and Drug Disposition	0142-2782	Life Sciences	12	
Computer Animation and Virtual Worlds	1546-4261	Physical Sciences	24	
Concurrency and Computation: Practice & Experience	1532-0626	Physical Sciences	24	
Contrast Media and Molecular Imaging	1555-4309	Physical Sciences	24	
European Law Journal	1351-5993	Social Sciences & Humanities	24	
European Transactions on Electrical Power	1430-144X	Physical Sciences	24	
Forest Pathology	1437-4781	Life Sciences	12	
Higher Education Quarterly	0951-5224	Social Sciences & Humanities	24	
Hippocampus	1050-9631	Life Sciences	12	
Infant and Child Development	1522-7227	Social Sciences & Humanities	24	
International Journal for Numerical Methods in Engineering	0029-5981	Physical Sciences	24	
International Journal of Adaptive Control and Signal Processing	0890-6327	Physical Sciences	24	

International Journal of Applied Linguistics	0802-6106	Social Sciences & Humanities	24	
International Journal of Osteoarchaeology	1047-482X	Life Sciences	12	
International Journal of Systematic Theology	1463-1652	Social Sciences & Humanities	24	
Journal of Advanced Nursing	0309-2402	Medicine	12	
Journal of Clinical Periodontology	0303-6979	Medicine	12	
Journal of Molecular Recognition	0952-3499	Physical Sciences	24	
Journal of Sociolinguistics	1360-6441	Social Sciences & Humanities	24	
Luminescence	1522-7235	Physical Sciences	24	
Marine Ecology	0173-9565	Life Sciences	24	
Modern Theology	0266-7177	Social Sciences & Humanities	24	
Particle and Particle Systems Characterization	0934-0866	Physical Sciences	24	
Polymers for Advanced Technologies	1042-7147	Physical Sciences	24	
River Research and Applications	1535-1459	Life Sciences	12	
Social Policy & Administration	0144-5596	Social Sciences & Humanities	24	
Zoo Biology	0733-3188	Life Sciences	12	

* Authors are recommended to refer to the **PEER Helpdesk** (http://peer.mpdl.mpg.de/helpdesk/wiki/embargoperiods) for an explanation of the embargo period and how this relates to their submissions to participating PEER repositories

PEER: Publisher submission Journal list by publisher

Publisher/ Journal	ISSN	Broad Classification	Embargo (months)	Language (if not Eng)
BMJ Publishing Group				
British Journal of Ophthalmology	0007-1161	Medicine	6	
Journal of Epidemiology and Community Health	0143-005X	Medicine	6	
Tobacco Control	0964-4563	Medicine	5	
EDP Sciences				
ESAIM: Probability and Statistics	1292-8100	Physical Sciences	12	French/ Eng
The European Physical Journal − Applied Physics	1286-0042	Physical Sciences	12	
Elsevier				
Annales Medico-Psychologiques	0003-4487	Medicine	18	French
Applied Thermal Engineering	1359-4311	Physical Sciences	24	
Astroparticle Physics	0927-6505	Physical Sciences	18	

Biochemical Pharmacology	0006-2952	Life Sciences	12	
Biochimica et Biophysica Acta (BBA) – Molecular Basis of Disease	0925-4439	Life Sciences	12	
Biophysical Chemistry	0301-4622	Physical Sciences	18	
Composites Science and Technology	0266-3538	Physical Sciences	18	
Computer Speech & Language	0885-2308	Physical Sciences	18	
European Journal of Mechanics − A/Solids	0997-7538	Physical Sciences	24	
Experimental and Toxicologic Pathology	0940-2993	Life Sciences	18	
Experimental Gerontology	0531-5565	Medicine	18	
Human Movement Science	0167-9457	Life Sciences	18	
Icarus	0019-1035	Physical Sciences	18	
International Journal of Impact Engineering	0734-743X	Physical Sciences	24	
International Journal of Non-Linear Mechanics	0020-7462	Physical Sciences	18	
Journal of Econometrics	0304-4076	Social Sciences & Humanities	36	
Journal of Economic Behavior & Organization	0167-2681	Social Sciences & Humanities	36	
Journal of Economic Dynamics & Control	0165-1889	Social Sciences & Humanities	36	
Journal of Experimental Social Psychology	0022-1031	Social Sciences & Humanities	36	
Journal of Geodynamics	0264-3707	Physical Sciences	18	
Journal of Physics and Chemistry of Solids	0022-3697	Physical Sciences	18	
Marine Environmental Research	0141-1136	Life Sciences	12	
Molecular and Cellular Endocrinology	0303-7207	Life Sciences	12	
Physics of the Earth and Planetary Interiors	0031-9201	Physical Sciences	24	
Pulmonary Pharmacology & Therapeutics	1094-5539	Medicine	18	
Speech Communication	0167-6393	Physical Sciences	18	
Statistics & Probability Letters	0167-7152	Physical Sciences	24	
Veterinary Microbiology	0378-1135	Medicine	18	
IOP Publishing				
Journal of Physics B: Atomic, Molecular and Optical Physics	0953-4075	Physical Sciences	12	
Journal of Physics D: Applied Physics	0022-3727	Physical Sciences	12	
Journal of Physics G: Nuclear and Particle Physics	0954-3899	Physical Sciences	12	
Nature Publishing Group				
Cell Death and Differentiation	1350-9047	Life Sciences	6	
European Journal of Clinical Nutrition	0954-3007	Medicine	6	
European Journal of Human Genetics	1018-4813	Life Sciences	6	
Molecular Psychiatry	1359-4184	Medicine	6	

Nature Immunology	1529-2908	Life Sciences	6	
Nature Neuroscience	1097-6256	Life Sciences	6	
Neuropsychopharmacology	0893-133X	Life Sciences	6	
Prostate Cancer and Prostatic Diseases	1365-7852	Medicine	6	
Oxford University Press				
International Journal of Epidemiology	0300-5771	Medicine	12	
Journal of Plankton Research	0142-7873	Life Sciences	12	
Portland Press				
Clinical Science	0143-5221	Medicine	12	
Springer				
Agriculture and Human Values	0889-048X	Social Sciences & Humanities	24	
Annals of Hematology	0939-5555	Medicine	12	
Breast Cancer Research and Treatment	0167-6806	Medicine	12	
Crime Law and Social Change	0925-4994	Social Sciences & Humanities	24	
European Child & Adolescent Psychiatry	1018-8827	Social Sciences & Humanities	24	
European Journal of Clinical Pharmacology	0031-6970	Life Sciences	12	
European Journal of Population	0168-6577	Social Sciences & Humanities	6	
European Journal of Wildlife Research	1612-4642	Life Sciences	12	
Formal Aspects of Computing	0934-5043	Physical Sciences	12	
Helgoland Marine Research	1438-387X	Physical Sciences	12	
Journal of Public Health	0943-1853	Social Sciences & Humanities	6	
Journal of Seismology	1383-4649	Physical Sciences	12	
Linguistics and Philosophy	0165-0157	Social Sciences & Humanities	24	
Review of World Economics	1610-2878	Social Sciences & Humanities	24	
Revue de Synthese	0035-1776	Social Sciences & Humanities	24	French
Taylor & Francis Group				
Aids Care	0954-0121	Life Sciences	12	
Applied Economics	0003-6846	Social Sciences & Humanities	18	
Avian Pathology	0307-9457	Life Sciences	12	
British Poultry Science	0007-1668	Life Sciences	12	
Communications in Statistics – Simulation and Computation	0361-0918	Physical Sciences	12	
Engineering Optimization	0305-215X	Physical Sciences	12	

Ethnic and Racial Studies	0141-9870	Social Sciences & Humanities	18	
Europe-Asia Studies	0966-8136	Social Sciences & Humanities	18	
Food Additives & Contaminants (Part A)	0265-203X	Life Sciences	12	
International Journal of Computer Integrated Manufacturing	0951-192X	Physical Sciences	12	
International Journal of Computer Mathematics	0020-7160	Physical Sciences	12	
International Journal of Production Research	0020-7543	Physical Sciences	12	
International Journal of Science Education	0950-0693	Social Sciences & Humanities	18	
Journal of Development Studies	0022-0388	Social Sciences & Humanities	18	
Molecular Physics	0026-8976	Physical Sciences	12	
Molecular Simulation	0892-7022	Physical Sciences	12	
Philosophical Magazine	1478-6435	Physical Sciences	12	
Psychology and Health	0887-0446	Social Sciences & Humanities	12	
Quantitative Finance	1469-7688	Social Sciences & Humanities	18	
Regional Studies	0034-3404	Social Sciences & Humanities	18	
Supramolecular Chemistry	1061-0278	Physical Sciences	12	
Technology Analysis & Strategic Management	0953-7325	Social Sciences & Humanities	18	
Wiley-Blackwell				
Alimentary Pharmacology & Therapeutics	0269-2813	Medicine	12	
Allergy	0105-4538	Medicine	12	
American Journal of Hematology	0361-8609	Medicine	12	
Bioethics	0269-9702	Social Sciences & Humanities	24	
Biotechnology Journal	1860-6768	Life Sciences	12	
British Journal of Haematology	0007-1048	Medicine	12	
Cell Biochemistry and Function	0263-6484	Life Sciences	12	
Clinical Endocrinology	0300-0664	Medicine	12	
Corporate Governance	0964-8410	Social Sciences & Humanities	24	
Developing World Bioethics	1471-8731	Social Sciences & Humanities	24	
Developmental Science	1363-755X	Social Sciences & Humanities	24	
Electrophoresis	0173-0835	Life Sciences	12	
Fuel Cells	1615-6846	Physical Sciences	24	
Global Change Biology	1354-1013	Life Sciences	24	

Haemophilia	1351-8216	Medicine	12	
Histopathology	0309-0167	Medicine	12	
Human Brain Mapping	1065-9471	Life Sciences	12	
Human Mutation	1059-7794	Life Sciences	12	
International Journal of Clinical Practice	1368-5031	Medicine	12	
Journal of Clinical Ultrasound	0091-2751	Medicine	12	
Journal of Community and Applied Social Psychology	1052-9284	Social Sciences & Humanities	24	
Journal of Medical Virology	0146-6615	Medicine	12	
Journal of Physical Organic Chemistry	0894-3230	Physical Sciences	24	
Molecular Microbiology	0950-382X	Life Sciences	12	
Oral Diseases	1354-523X	Medicine	12	
Pediatric Anesthesia	1155-5645	Medicine	12	
Pediatric Pulmonology	8755-6863	Medicine	12	
Phytotherapy Research	0951-418X	Life Sciences	12	
Social Development	0961-205X	Social Sciences & Humanities	24	
ZAAC − Zeitschrift für anorganische und allgemeine Chemie / Journal of Inorganic and General Chemistry	0044-2313	Physical Sciences	24	German / English

Appendix B. Technical specifications for CSV metadata provision

The CSV file must conform with the following specifications:

- Filename not important, but extension must be '.csv'
- UTF8 encoding
- Quote character **"**
- Separation character **,**
- End-of-line ***\n***
- Field names included in the first line
- Field names must be among :

Textual column title	Comment
author_country	country code ISO 3166-1-A2
author_firstname	
author_middle	
author_lastname	
author_email	
affiliation_institution	
affiliation_department	
pubdate	ISO 8601
article_title	
journal_title	
publisher_article_id	
doi	
abstract	
issn	
volume	
issue	
fpage	
lpage	
subject	
Lang	ISO 639-3
embargo	In months

We insist that, except for 'publisher_article_id' and 'doi' which are used for linking both passes, there is no overlapping between the metadata sets of both passes. Metadata submitted twice will not be updated.

Appendix C. The SWORD protocol

1 Introduction

In the PEER project, selected stage-2 material from publishers is being transferred to or deposited into the PEER Depot after which the content is being transferred from the depot to multiple, publicly available repositories.

The stage-2 material will be transferred in a Submission Information Package (SIP) containing the full-text publication, metadata and the complementary stage-2 source files. The SWORD AtomPub profile contains specific features that allows for an application-level deposit of material into repositories.

The PEER information model can be mapped onto the OAIS Reference Model and the DRIVER object model for Enhanced Publications.

Implementers may set up their own server conforming to these guidelines using one of repository specific implementations available from SourceForge, or write their own custom implementation either using the generic Java library, also available from SourceForge, begin their implementation from scratch.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

It is assumed that the reader of this document has knowledge of the PEER D2.1 report[1], SWORD profile v1.3[2] , the OAIS[3] Reference Model[4] and the DRIVER[5] II Enhanced Publication object model and Functionalities[6].

1.1 SWORD overview

The SWORD AtomPub Profile is an application profile of the Atom Publishing Protocol (APP) (RFC 5023)[7] that contains specific features that allows for an application-level deposit of material into repositories.

The APP is based on the HTTP transfer of Atom-formatted representations. It is easy to think of APP as a way of publishing just Atom Syndication Format feeds. While it is true that APP provides the means to publish Atom Syndication Format Entries to collections (such as blogs), it also provides a mechanism for the publishing of binary formatted data called Media Resources in APP context (Internet Engineering Task Force 2007). While in the blog scenario this mechanism may be used to add attachments to a blog post i.e. images, audio, video, documents), SWORD exploits this for the publishing (or deposit) of material into repositories, usually in some form of content packaging in which data and descriptive metadata are being held together in one container (see Figure 8).

---

1 PEER D2.1 *Draft report on log file harvesting systems and manuscript deposit procedures for publishers and repository managers*, http://www.peerproject.eu/reports/

2 Allinson, J et al 2008, *SWORD AtomPub Profile version 1.3*, viewed 25 March 2009 http://www.swordapp.org/docs/sword-profile-1.3.html

3 Open Archival Information System.

4 Consultative Committee for Space Data Systems 2002, *OAIS Reference Model* http://public.ccsds.org/publications/archive/650x0b1.pdf

5 Digital Repository Infrastructure Vision for the European Region.

6 Verhaar, P & Place, T 2008, *Report on Object Models and Functionalities*, DRIVER II D4.2.

7 Internet Engineering Task Force 2007, *The Atom Publication Protocol*, RFC 5023, Internet Engineering Task Force, http://tools.ietf.org/html/rfc5023

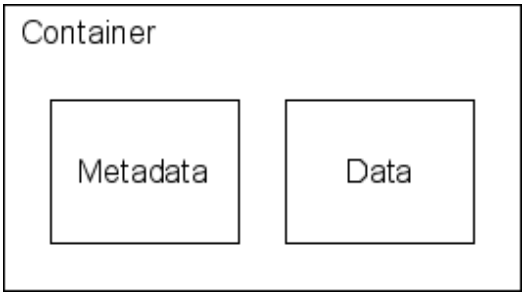

Figure 8: Content Package or Container

An example of an implementation of such a container would be a ZIP-file containing a full-text manuscript in the PDF/A-1 format and descriptive metadata in the TEI-XML format.

The container is being submitted by a client to a SWORD interface service (server) as a bit stream using a HTTP POST request consisting of a header containing information about authorisation and the bit stream (type and format of the container) in order for the server to be able to interpret the bit stream properly, and a body part containing the bit stream itself (see Figure 9). Upon reception, the server sends a HTTP response back to the client – again consisting of a header and a body part – with the header containing a HTTP status code indicating a success or failure of the attempted deposit according to regular HTTP semantics, and a response document containing additional APP/SWORD specific information about the deposit being made.

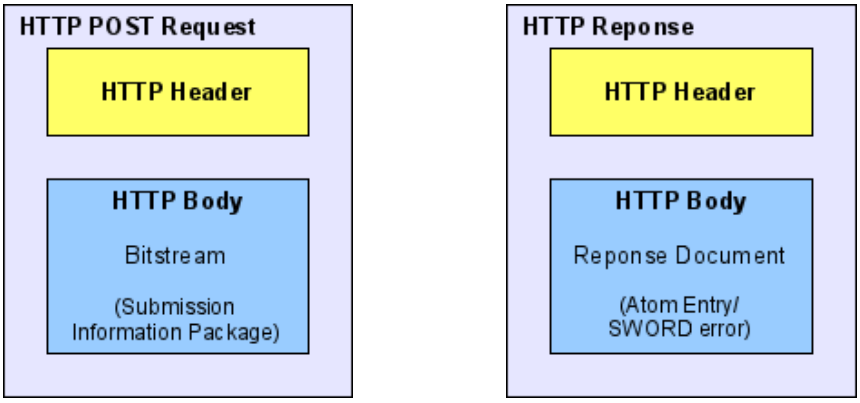

Figure 9: HTTP request and response structure in the SWORD context

1.2 Use of SWORD in PEER

In the PEER workflow there are two scenarios of deposits into the PEER repositories specified: deposit made by PEER and deposit made by authors (see Figure 10)

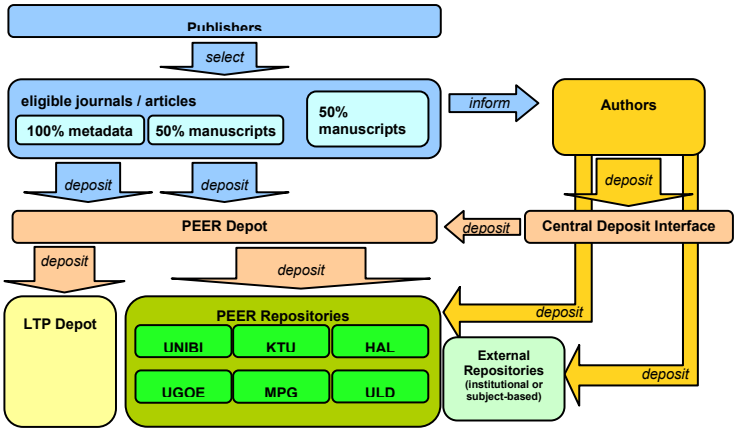

Figure 10: PEER Workflow

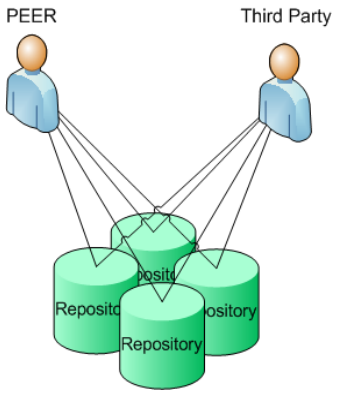

Figure 11: Deposit situation

This results in an n:n-relation between repositories and deposit sources either the PEER Depot or third party services operated by an author (see Figure 11). To prevent multiple tailored solutions and implementations it is important to define a standard process for the deposit of material into repositories.

The processes may be categorised into two types of mechanisms: **push and pull**. An example of the **pull** mechanism is the KB's mechanism of the e-depot harvesting repositories through OAI-PMH and pulling content using a webclient (see Figure 12) which downloads the objects specified in the location entries in the metadata.

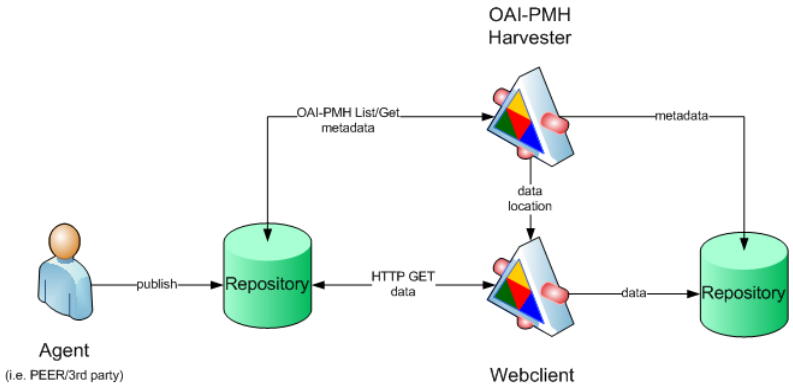

Figure 12: OAI-PMH data harvest

An example of the **push** mechanism is the SWORD deposit mechanism where the data is being pushed by an agent (i.e. a webservice or desktop application representing a user) to the SWORD interface of a repository which then accepts or rejects the deposit (see Figure 13).

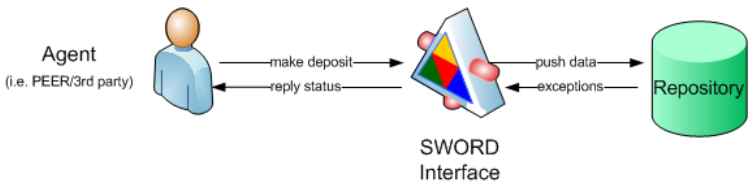

Figure 13: SWORD data deposit

Finally, a third, hybrid mechanism can be created by setting up an FTP server to which deposits can be uploaded (pushed) by an agent. A repository may then pull the FTP content which is then being pulled into the repository (see Figure 14).

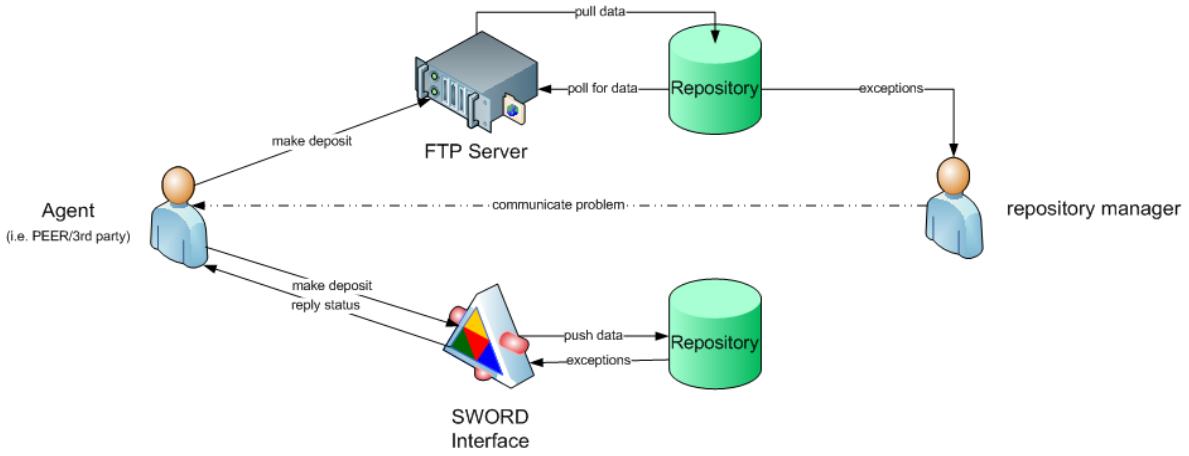

Figure 14: SWORD versus FTP

A disadvantage of this mechanism is that this only provides direct feedback to the agent about status of the upload, not of the status of the actual deposit into the repository. This may lead to the situation when an agent successfully uploads data to the FTP server, but the data is being rejected by the repository afterwards because it does not adhere to rules the repository enforces on its contents without the agent being informed about this rejection – something that is not the case when using SWORD.

Figure 15 provides a schematic overview of the use of SWORD in the PEER deposit scenario. Here a publisher transfers manuscripts and metadata into the PEER Depot where the manuscripts and metadata are being converted and crosswalked to the formats specified for the PEER deposit process. The converted and crosswalked manuscripts and metadata are then being packaged into a container and sent to the SWORD interface service of a repository where the contents are being unpacked from the container. Upon reception these MAY be converted and crosswalked into an internal storage format before they are being archived into the repository.

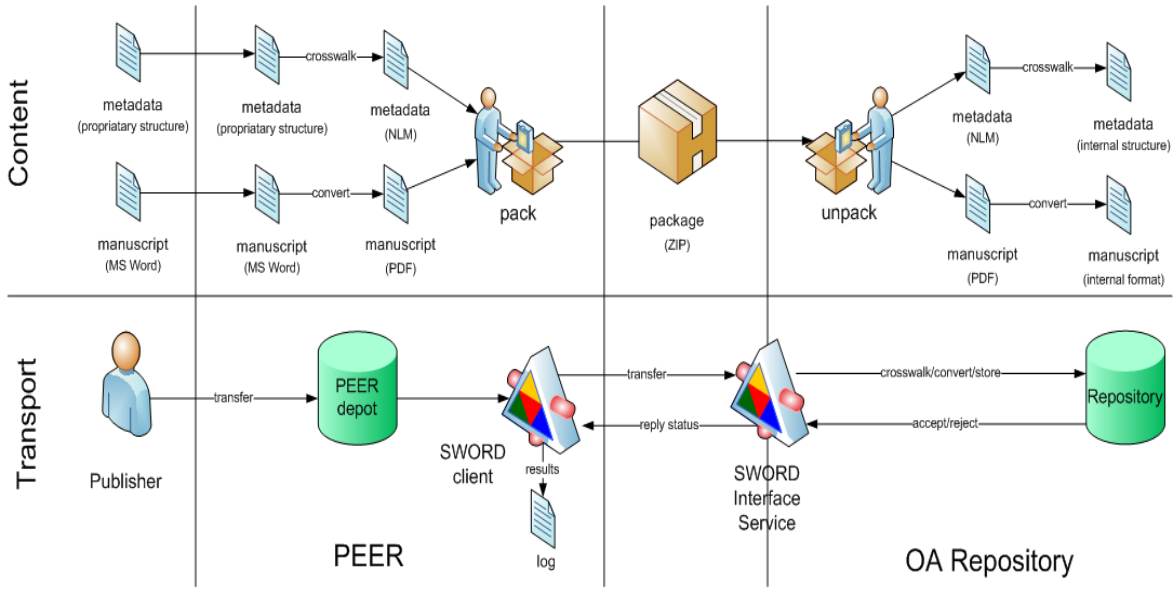

Figure 15: SWORD use in PEER for PEER Depot

2 Use of SWORD features

2.1 About this section

This section will describe the use of the SWORD profile in the context of the PEER project. The contents are organised according and supplementary to the document SWORD Atom Pub Profile version 1.3 part A. If a SWORD profile section or feature is omitted, implementations MUST behave as defined in SWORD profile.

2.2 Package Support

The PEER Submission Information Package (SIP) MAY be expressed using (a combination of) different formats (i.e. XML containers or RFC 1951 compliant ZIP archives) and/or serialised using different structural models (i.e. DIDL, METS, ORE, TEI, NLM, MODS, DC). The mappings between the SIP, its components and the formats and structures will be defined and expressed using specialised application profiles developed in the PEER context.

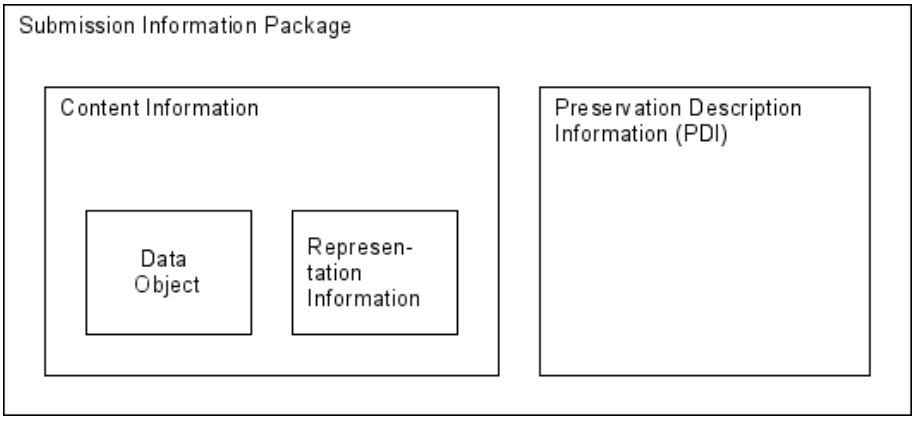

Figure 16: Submission Information Package structure

The SWORD profile offers the possibility to enumerate multiple packaging formats in the Service Document and supply a Quality Value attribute indicating a preference and level of support for a designated package format.

2.2.1 Package support in Service Description

The server MAY support multiple packaging formats with varying quality values according to the support of the PEER Submission Information Package (SIP).

The server MUST support at least one package format with Quality Value “1.0”, indicating full support where all components supplied within the SIP will be processed and understood when using the designated package format.

All supported formats MUST be listed in the Service Document.

All formats listed in the Service Document MUST have a Quality Value attribute assigned.

The value used in the <sword:accepted Packaging> element MUST NOT overload any values enumerated in the SWORD Content Package Types.

The server MAY use the <sword:service> element in the Service Document to indicate the existence of other service interfaces supporting additional package formats.

The server SHOULD NOT accept a specific package format across multiple interfaces with different levels of support as indicated by the Quality Value attribute in the Service Document.

2.2.2 Package Support during Resource Creation

If a server receives a POST request with a format that is not listed as an accepted format in the Service Document, the server MUST reject the package by returning an HTTP status code of 415 (unsupported media type).

2.2.3 Package description in entry documents

When describing packaged resources in Media Entry documents, the server SHOULD add sword:packaging elements to the entry.

2.3 Mediated Deposit

The following paragraph is considered informative, but is included for clarity in the use of the SWORD profile outside the PEER project.

The PEER workflow offers two ways a manuscript can be deposited into one of the publicly available PEER repositories: either by publisher deposit (through the PEER Depot) or by author deposit (where the publisher informs the author who deposits his/her article(s) via the depot interface /the PEER Depot at the actual publicly available repositories).

For the author deposit, the author MAY make the deposit by proxy through a web service (i.e. by filling in a form to provide the metadata and upload a file containing the full-text material) after which the web service is making the actual deposit. The web service MAY not be used for the PEER project exclusively in which case the web service MAY use its own credentials to authenticate at the server (at the repository side).

Figure 17 depicts an example of the use of this mechanism in the PEER context. Note that the greyed out parts of the figure are considered outside the scope of the PEER project.

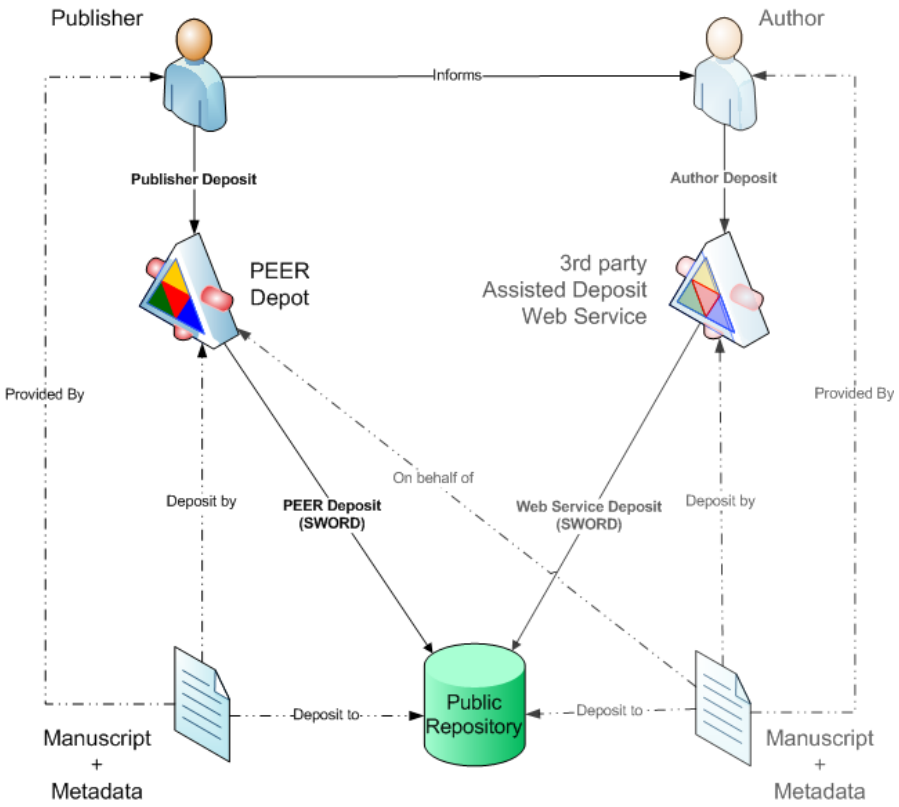

Figure 17: PEER deposit workflow

It is recognised that the repository MAY want to keep track of data that is being deposited within the PEER context by creating a single user account to the PEER Depot. This then covers the publisher deposit workflow, but does not provide for a solution for the case of author deposit through another web service which MAY use different credentials.

A possible solution MAY be the use of mediated deposit where a client authenticates using its assigned credentials on behalf of another known user (e.g. a web service authenticates using its own credentials and makes the deposit on behalf of the PEER user which is used by the PEER Depot).

This method MAY also be used to authenticate on behalf of other users (i.e. authors, librarians, data stewards, research assistants, etc.) that already have a valid user account at the repository.

The use of mediated deposit is considered OPTIONAL and is currently not implemented in the application of the SWORD profile within the PEER project.

2.3.1 Mediation in Service Description

Servers supporting mediated deposit MUST indicate this by including a SWORD:mediation element with a value of “true” in the Service Document as defined in the SWORD profile version 1.3 section 2.1.

For servers that do not include a SWORD mediation element in the Service Document, a default value of “no” SHOULD be assumed by clients.

2.4 Auto-discovery

AtomPub makes no recommendations on the discovery of Service Documents.

The SWORD profile states that it is RECOMMENDED that server implementations use an <html:link rel="sword" href="[Service Document URL]"/> element in the head of a relevant HTML document to assist with service discovery.

In addition, it is RECOMMENDED to also include an <atom:link rel="sword" type="application/atomsvc+xml" href="[Service Document URL]"/> element in relevant response documents such as Error Documents.

2.5 Nested Service Descriptions

Nested Service Descriptions MAY be used to specify alternative collections for both organisational (i.e. generic collection with a nested PEER specific collection) and technical purposes (i.e. a specific interface or service instance to cater for specific types of content packaging).

3 Use of APP features

The contents of the following section are organised according and supplementary to the document SWORD Atom Pub Profile version 1.3 part B. If a SWORD profile section or feature is omitted, implementations MUST behave as defined in SWORD profile.

3.1 Securing the Atom Publishing Protocol

The SWORD profile states servers SHOULD support the use of HTTP Basic Authentication over TLS. Because from a trust perspective it is important to confirm the identity of the PEER Depot during the deposit proces, this statement is considered insufficient for the purposes of the PEER project.Therefor this requirement has been restated as follows:

Servers implementing SWORD MUST support HTTP Basic Authentication (RFC 2617) over TLS (RFC 2818).

3.2 Creating and Editing Resources

When depositing resources using SWORD, resources are created by a server when a client makes an HTTP POST request with the resource in the HTTP request body. If the deposit is made successfully, the server then gives a HTTP reponse with the HTTP 201 Status code in the header of the response indicating the resource has been successfully created at the repository side.

Servers returning a HTTP 201 status code after a deposit MUST preserve the resource deposited.

Clients receiving a HTTP 201 status code MUST consider the resource deposited as being accepted for storage by the repository.

3.2.1 Asynchronous treatment of resources

It MAY however be the case that the repository implements an additional asynchronous validation process after which a resource MAY or MAY NOT be accepted. This for instance is the case when a repository uses an intermediate repository where resources deposited through the SWORD interface are temporarily stored, after which they will be moved to a final location within the repository when they are properly validated by a repository manager. When a resource is then being rejected by the repository during the validation process after the server has sent an HTTP 201 response to the client, the situation MAY arise where the client considers the resource as being successfully deposited into the repository, while in fact the resource is NOT being stored into the repository. This situation is viewed as undesirable.

Servers implementing an asynchronous validation process MUST return an HTTP 202 Accept response code indicating the request has been accepted for processing, but the processing has not been completed.

Clients receiving a HTTP 202 status code upon deposit of a resource MUST consider the resource deposited as NOT being stored into the repository.

RFC2616 states that there is no facility for the re-sending of status codes. Therefore, a client will not receive a notification of the outcome of the processing carried out by the server. In order to allow clients to retrieve the outcome of the deposit, the sword:treatment element MAY contain the status of the processing of the deposited resource.

Servers implementing HTTP 202 status codes MUST supply a permanent link to the Atom Entry document of the response.

Servers implementing HTTP 202 status codes MUST update the sword:treatment element of the Atom Entry document of the resource with the status of the processing of the deposited resource.

Client SHOULD implement a mechanism to confirm the successful deposit by periodically checking back at the server with an HTTP GET request to the permanent link supplied by the server, in order to check the contents of the sword:treatment element of the Atom Entry describing the deposited resource when a HTTP 202 status code has been received upon deposit.

4 PEER Object Model

In order to future proof agreements and guidelines for a technical model, it is important to detach the technical implementation from the abstract object and information model. Furthermore, it is important to keep this abstract model aligned with other developments in the area the model will be used in. For PEER, there are two of such developments:

- OAIS Reference Model for its use by the KB
- DRIVER object model for Enhanced Publications for its use in DRIVER context

In PEER, manuscripts and metadata will be transferred between authors, publishers, the PEER Depot, Open Access repositories and an LTP Depot exploited by the KB.

This results in a PEER object consisting of a manuscript object which is being described by one or more metadata objects (see Figure 18).

The D2.1 report provides an exhaustive metadata field set to be used in the PEER project (see Table 5, below).

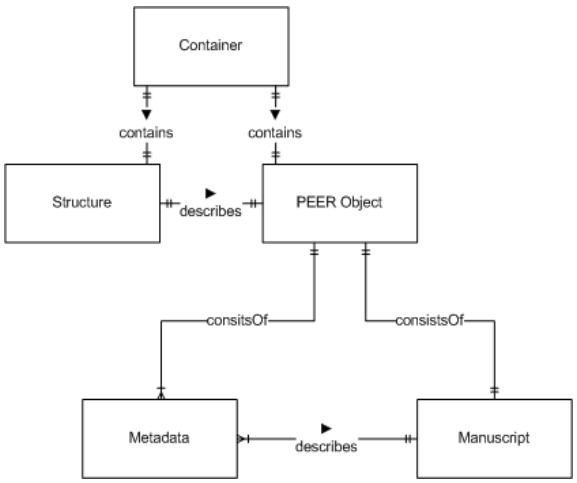

Figure 18: PEER Object model ERD

Field Name	Semantics	Syntax
Title	Article Title	
Creator	Corresponding author's name	Last name, first name
AuthorEmail	Corresponding author's e-mail address	
Description	Abstract	
Date	Date of publication	ISO 8601:2004 ; yyyy-MM-dd
Identifier	DOI of published article	
Coverage	Geographic location of the contributing Author	ISO 3166-1-A2
Journal	Journal title	
Affiliation	multi-tier organisation list	Country, Organisation, Laboratory
ISSN		
Volume		
Issue		
Page		
Type	Semantic type of the publication	info:eu-repo/semantics/article info:eu-repo/semantics/acceptedVersion defaults to article.
Subject	Subject headings; Scientific classification (defaults to what is provided in the PEER Journal tables[1])	
Language	Language of the publication	ISO 639-3 (defaults to 'eng')
Embargo	Embargo Period (defaults to what is provided in the PEER Journal tables)	

Table 5: PEER information model

1 See Appendix A.

For deposit the PEER object will be packaged into a container. The OAIS reference model specifies the Submission Information Package (SIP) as a specialised Information Package (IP) – which is used by the KB in the e-depot – for submission purposes (see Figure 19).

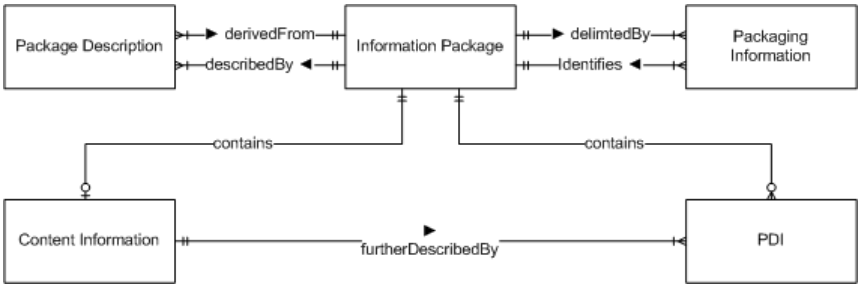

Figure 19: OAIS Information Package ERD

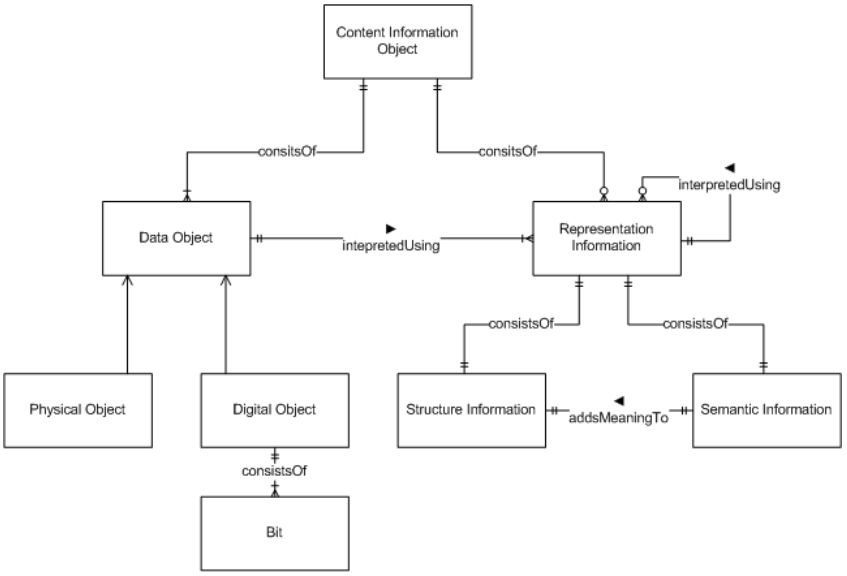

Figure 20: OAIS Content Information Object ERD

An IP consists of content information that is being described by Package Description Information (PDI).

The content information object is defined as a data object (i.e. PDF file) interpreted using representation information (i.e. mime-type, encoding version, etc.) (see Figure 20). Note the structure information being a part of the representation information.

The PDI (see Figure 21) contains Reference Information (i.e. bibliographic descriptions and persistent identifiers), Provenance Information (i.e. information about the conversion process), Context Information (i.e. reference to the research project a publication is based on) and Fixity Information (i.e. a checksum).

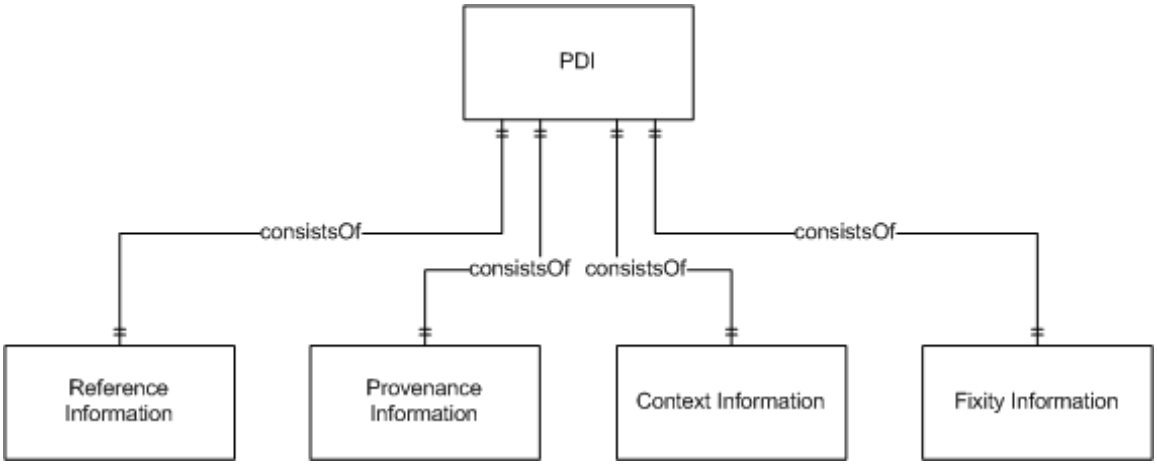

Figure 21: OAIS Package Description Information ERD

The PEER information model can be mapped onto the OAIS Reference Model as depicted in Figure 22. Here the structure object containing the structural information is being added to the PEER object model.

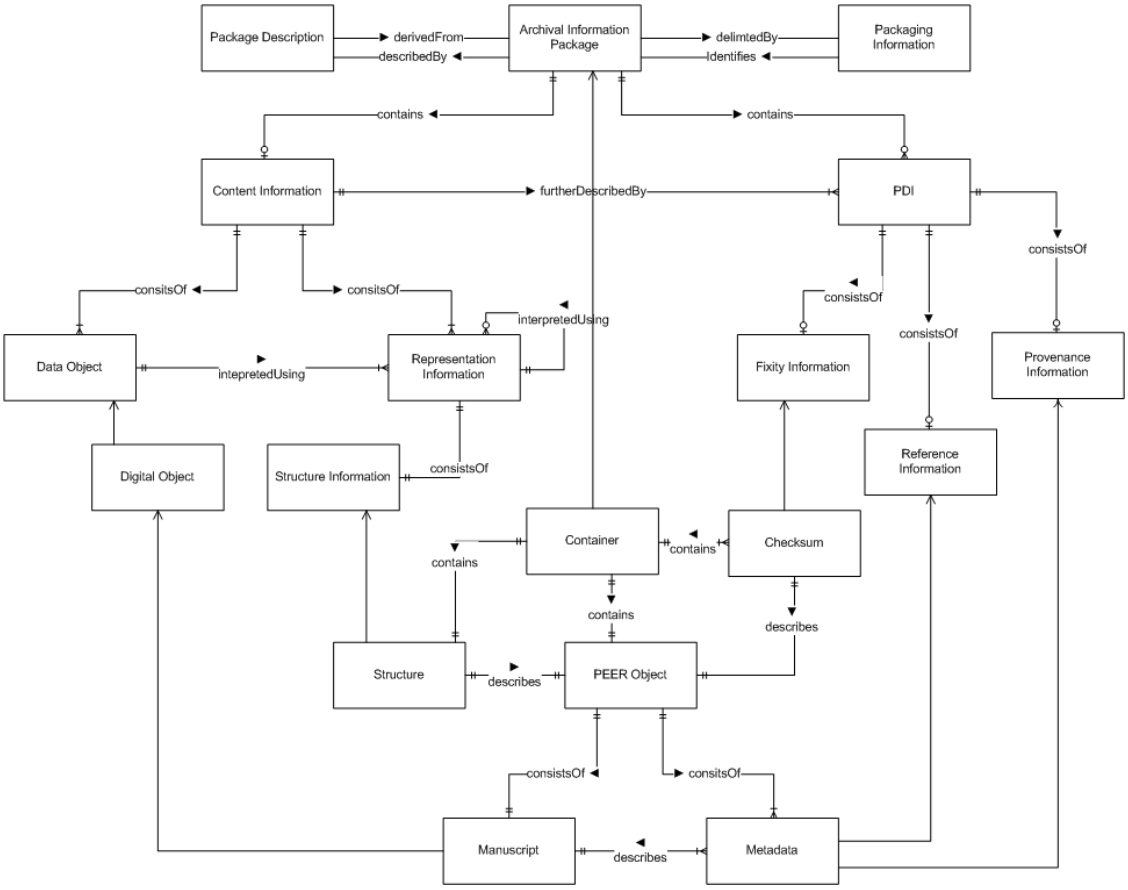

Figure 22: OAIS Reference Model-PEER Information Mapping

Figure 23 depicts a technical mapping of the PEER object model. The structure is expressed using the ORE abstract data model which is serialised as an Atom feed. This Atom feed is to be contained in an XML file. The metadata is serialised in TEI, again contained in an XML file. The manuscript is encoding in PDF/A, contained in a PDF-file.

The XML file containing the ORE Atom feed, the XML file containing the TEI document and the PDF file containing the manuscript are then being packaged in a compliant ZIP-file. Upon deposit using SWORD, the ZIP-file is being placed into the body of the HTTP POST request. The Header contains an MD5 checksum and MAY contain authorisation information (see Figure 24).

The HTTP POST request is then being sent to the SWORD Interface Service as described in paragraph 1.2.

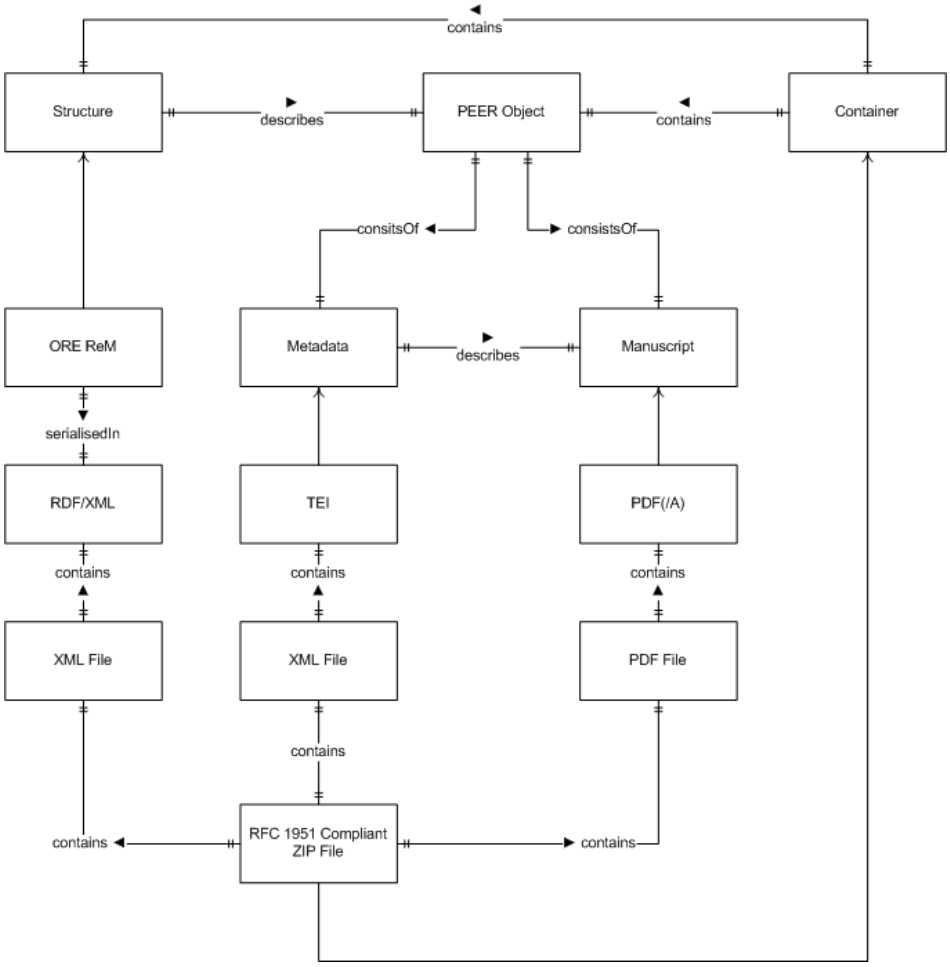

Figure 23: Technical Mapping of the PEER model

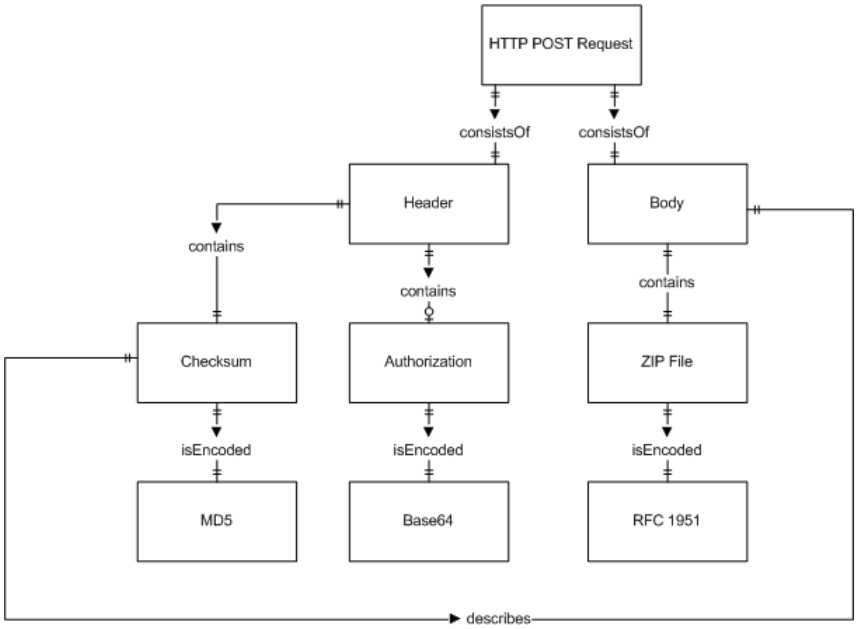

Figure 24: HTTP Mapping of the Technical Model

5 Implications on Repository level

5.1 Overview of the technical process

Generally speaking, the technical process of deposit can be broken in sequential order to the serialisation and deposit request sub-processes on the client side (i.e. the PEER Depot) and the de-serialisation, response and store sub-processes on the repository side (i.e. publicly available repository).

5.1.1 Serialisation – Client side

The serialisation processes involve the serialisation of the metadata from an internal storage to a specific (agreed upon standard) metadata field set and structure (i.e. DC) and the packaging of the metadata and object file(s) into a content package (i.e. MPEG-21 DIDL XML containers or compliant ZIP archives) which MAY include adding a manifest describing contents and their correlation (i.e. the relation between an XML file containing the descriptive metadata of a full-text publication and a PDF-file containing the actual full-text publication) to a bit stream.

5.1.2 Deposit Request – Client side

The deposit request process includes a client posting the data to a service (i.e. the HTTP POST request in SWORD) and the server receiving the data and placing it into a temporary storage (either in memory or on disk).

5.1.3 De-serialisation – Repository side

In the de-serialisation process the receiving server tries to interpret (decode) the bit stream again, in essence validating the contents. This MAY include the unpacking (when using ZIP archives) or decoding (when using XML containers) of the bit stream to be able to interpret the individual contents. It MAY also include the mapping or crosswalking of the metadata structure to an internal (proprietary) metadata field set and/or structure. This process MAY not necessarily be taking place in the actual interface service; it MAY include the sending of the bit stream to an internal storage service which then indicates a success or failure to the deposit service.

5.1.4 Response – Repository side

After the contents are being de-serialised successfully and the server confirms the contents of the received bitstream, the server MUST reply its status to the client (when using SWORD this means reporting the appropriate HTTP status code and correct Atom/SWORD response document). If the deserialisation process has failed for whatever reason, the server SHOULD reject the deposit request to indicate an unsuccessful deposit to the client, or accept the deposit with an HTTP 4xx status code and appropriate exception message indicating a partial successful deposit.

5.1.5 Store – Repository side

The final step of the deposit process includes the storage of the received (meta)data into the internal (meta)data store. This part of the process is implementation specific and considered outside the scope of this document.

5.2 Functional Requirements

A repository implementing a SWORD interface service in the PEER context MUST be able to:

- Authenticate a user
- Receive, process and respond to an HTTP POST request as specified in this document
- Interpret and store a PEER Submission Information Package as specified by the PEER project

5.3 Implementation Steps

The implementation steps can be broken down into the implementation and exposure of the web service to the outside world and interface with the repository on the inside.

Depending on specific needs, an implementer of the SWORD profile may either choose to make an implementation by using one of the repository specific implementations available for DSpace, ePrints and Fedora on Sourceforge[1] or to write a custom SWORD server implementation (optionally by using the generic Java library also available from Sourceforge).

For the repository specific option please refer to the documentation provided with the designated packages.

The second option either involves writing a service from scratch or use the source code available from the SWORD Java library. This library contains ready to implement code for writing servers and clients.

In addition to creating the web service which behaves according to the guidelines specified in this document, special attention should be paid to the creation of crosswalk rules to map the expression(s) of the PEER SIP to the internal repository data structure and semantics.

5.4 Common implementation faults

In the initial testing phase, a number of common implementation faults have been identified. This paragraph will provide a brief overview of these findings and provide guidelines in order to avoid these mishaps.

---

1 SWORD Project, *SourceForge.net: SWORD − Project Web Hosting − Open Source Software*, viewed on 25 March 2009, http://sword-app.sourceforge.net/

5.4.1 Mandatory fields for Atom Entry

A number of issues have been identified with regard to the contents of the Atom Entry within the service response documents.

RFC 5023 (Atom Publishing Protocol) states:

> *Implementers are asked to note that [RFC4287] specifies that Atom Entries MUST contain an atom:summary element. Thus, upon successful creation of a Media Link Entry, a server MAY choose to populate the* ***atom:summary*** *element (as well as any other mandatory elements such as* ***atom:id****,* ***atom:author****, and* ***atom:title****) with content derived from the POSTed entity or from any other source. A server might not allow a client to modify the server-selected values for these elements.*

5.4.2 Field Semantics

All contents of the Atom Entry document are to be interpreted as within the SWORD context. For example, atom:author states the author of the transaction (i.e. authenticated user), not that of the contents being transported (i.e. author of the publication).

5.4.3 Atom:Id

The contents of the atom:id field MUST be encoded using IRI (International Resource Identifier) as defined in RFC3987.

Valid example:

```
<atom:id>http://hdl.handle.net/2437.2/20</atom:id>
```

Invalid example:

```
<atom:id>1234</atom:id>
```

Locations

When the server has processed the submission information package containing the fulltext and metadata files, a number of locations may be identified:

1. Location of the Media Link Entry
2. Location of the original submission information package
3. Location of the fulltext
4. Location of the metadata

   …

SWORD APP Profile v1.3 states:

> *The Location element of the HTTP header response MUST contain the URI of the Media Link Entry, as defined in ATOMPUB. The Media Link Entry URI MUST dereference, and MUST contain an atom:content element with a src attribute containing a URI.*

Example:

```
HTTP/1.1 201 Created
Date: Mon, 18 August 2008 14:27:11 GMT
Content-Length: nnn
Content-Type: application/atom+xml; charset="utf-8"
Location: http://www.myrepository.org/geo/atom/my_deposit.atom
```

```
<entry ...>
  <title>My Deposit</title>
  <id>info:something:1</id>
  <updated>2008-08-18T14:27:08Z</updated>
  <author><name>jbloggs</name></author>
  <summary type="text">A summary</summary>
  ...
  <content type="application/zip"
    src="http://www.myrepository.ac.uk/geo/deposit1.zip"/>
  <sword:packaging>http://purl.org/net/sword-types/tei/peer</sword:packaging>
  <link rel="edit"
    href="http://www.myrepository.org/geo/atom/my_deposit.atom" />
</entry>
```

The Server MUST use the correct MIME-type reference in the type attribute of the atom:content element in the Media Link Entry.

The Media Link Entry MUST contain an atom:element with a src attribute containing the location of the fulltext. This is used in the feedback e-mail to the author.

The Media Link Entry MAY contain an atom:element with a src attribute containing the location of the original submission information package, the raw metadata, ORE-ReM, jump-off page, etc.

Appendix D. Peer Author Deposit interface specification

Interface: http://peer.mpdl.mpg.de/deposit

Summary

Authors are invited to self-deposit publications to the PEER repositories.

Actors

Corresponding author

Flow of Events

1. A user chooses to deposit his/her publication to the PEER Depot.
2. The user can enter basic metadata by using a webform (Note: journal name is provided from a list)
3. The user can upload a PDF file.
   3.1. The system checks the file mimetype and gives an error message, if the file is not recognised as application/pdf.
4. The user needs to fill out a text shown on an image to avoid spamming (ReCAPTCHA[1] mechanism).
5. The user finalises the submission by submitting the form.
6. The webform performs a simple validation.
   6.1. The webform content is validated successfully:
      6.1.1. The system shows a confirmation message.
   6.2. The webform content is validated unsuccessfully:
      6.2.1. The system informs the user on missing/not populated mandatory fields, or wrong image recognition and asks the user to correct the entries and re-submit the form. The use case ends unsuccessfully.
7. The system packs the metadata and the PDF file into an archive and saves the content to a dedicated directory on the server (see Processing and Deposit of publications).
8. The use case ends successfully.

Additional information on depositing interface[2]

- All data will be deposited to all participating repositories.
- The user must select the journal name from a predefined list of journals.
- Metadata will be provided in TEI-XML format.
- Deposit interface will pack the metadata and provided PDF file into an archive.
- The archived data will be posted to the Peer Depot via ftps[3] protocol.

---

1 http://en.wikipedia.org/wiki/ReCAPTCHA
2 More details can be found at http://colab.mpdl.mpg.de/mediawiki/Peer:_Author_Deposit
3 http://en.wikipedia.org/wiki/FTPS

Appendix E. Alternate author deposit workflow scenarios

Scenario 1: Author deposit in repositories, after registration at repositories

- Author receives notification of acceptance, including invitation from publisher to self-archive stage-2 article. At this stage publication date is undetermined, and embargo period is unknown.
- Author follows invitation to access further details via PEER Helpdesk.
- Author selects a participating repository for deposit or enters an URL of additional/ alternative repository of choice.
- Author accesses the participating repository's website.
- Author registers[1] at the repository in order to be able to make deposit. Authors, including those not affiliated to the repository host institution, are authorised for deposit.
- Author deposits his/her data in the repository.
- Author receives a notification of successful deposit.
- Embargo management is handled by the repository.
- Repositories transfer their usage log files to the Usage Research Team.

Problems:

- Authors can be authenticated and authorised only if they access from the repository's host institution network (e.g. university network).
- Embargo management requires automated matching process at repositories: match author submitted article with metadata from publishers (transferred by PEER Depot) and overwrite author's metadata with publisher's metadata. Elements of manual checking may occur (e.g. special characters).
- Random deposit by non-PEER authors difficult to monitor, as there is no possibility to identify if potential deposits come by following a PEER invitation or not.

1 The author registration may vary from one repository to another e.g. sending an e-mail to the repository managers or fill-in a web form for requesting the privileges. For purpose of simplification of the workflow analysis these details are not further depicted.

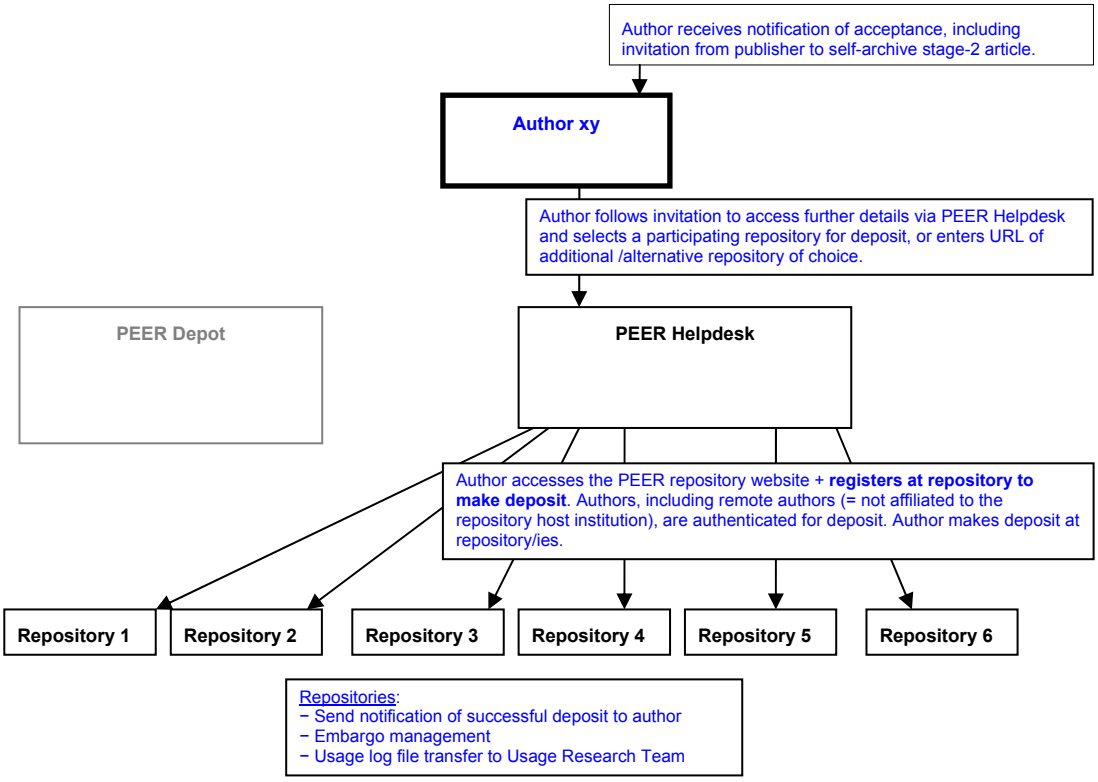

Figure 25: Scenario 1

Scenario 2: Central author registration at PEER Helpdesk. Author gets redirected to repositories for authentication under generic PEER account and deposit

- Author receives notification of acceptance, including invitation from publisher to self-archive stage-2 article. At this stage publication date is undetermined, and embargo period is unknown.
- Author follows invitation to access further details via PEER Helpdesk.
- *Author registers at the PEER Helpdesk.*
- Author selects a participating repository for deposit.
- Author enters URL of additional /alternative repository of choice.
- *Author is redirected to the chosen repository to make deposit.*
- *Author is authenticated under generic PEER guest account.*
- Author deposits his/her data in the repository.
- Author receives a notification of successful deposit.
- Embargo management is handled by the repository.
- Repositories transfer their usage log files to the Usage research team.

Problems:

- Setting up a central author registration application means an additional effort of funding and staff resources.

- Central registration is dependent upon authentication of generic PEER account at repositories.
- Embargo management requires automated matching process at repositories: match author submitted article with metadata from publishers (transferred by PEER Depot) and overwrite author's metadata with publisher's metadata. Elements of manual checking may occur (e.g. special characters).

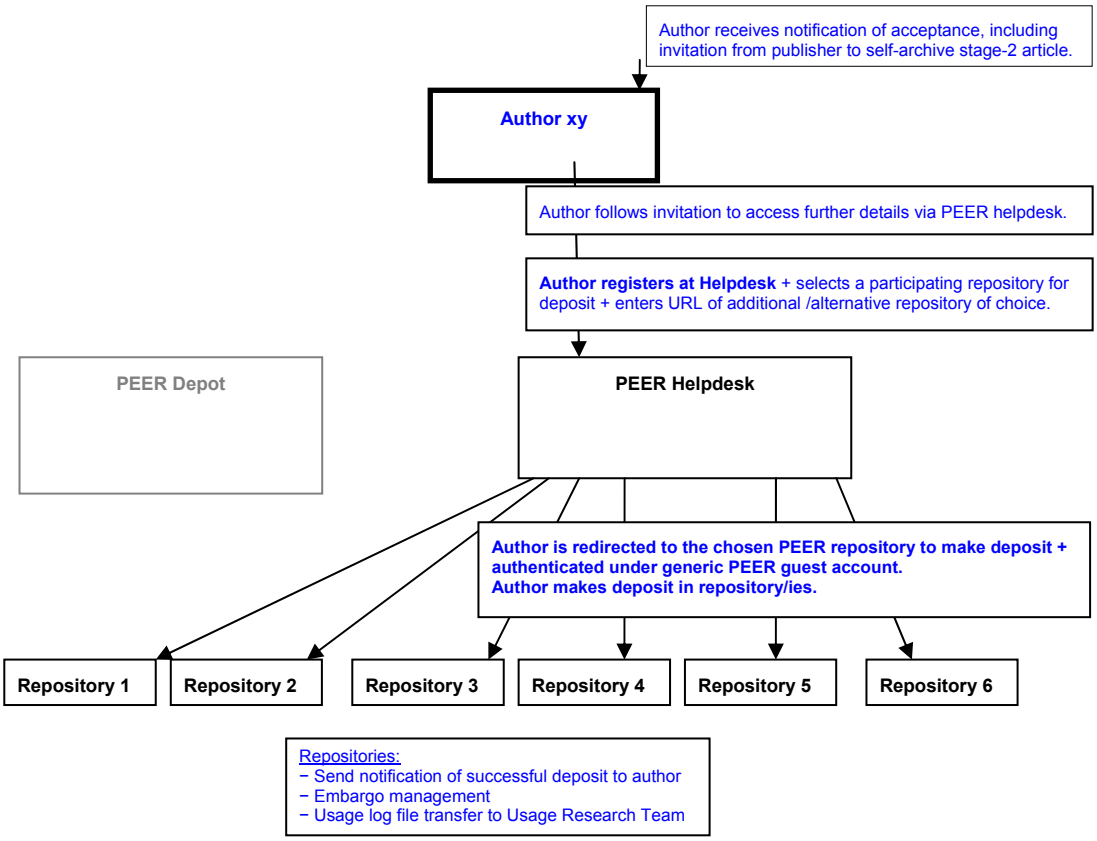

Figure 26: Scenario 2

It is possible to combine the central registration with a central deposit with the authors uploading their articles directly into the PEER Depot. Author deposits will in this scenario be handled like publisher deposits, with distribution to all participating repositories, and where duplicate filtering and embargo management will also happen at the PEER Depot.

The workflow for central author deposits is described in scenarios 3 & 4 below.

Scenario 3: Central registration at PEER Helpdesk, central deposit at PEER Depot via Helpdesk

- Author receives notification of acceptance, including invitation from publisher to self-archive stage-2 article. At this stage publication date is undetermined, and embargo period is unknown.
- Author follows invitation to access further details via PEER Helpdesk.
- Author registers at the PEER Helpdesk.
- *Author is redirected to a deposit interface on the PEER Depot.*

- *Author deposits his/her data in the PEER Depot.*
- Author receives notification of successful deposit.
- Embargo management at the PEER Depot, as per publisher deposits.
- Distribution of author deposits to all participating repositories, as per publisher deposits.
- Repositories transfer their usage log files to the Usage research team.

Problems:

- Central deposit seen as interference with the 'natural way' of author deposit – referred to and approved by Executive.
- No communication between authors and repositories.

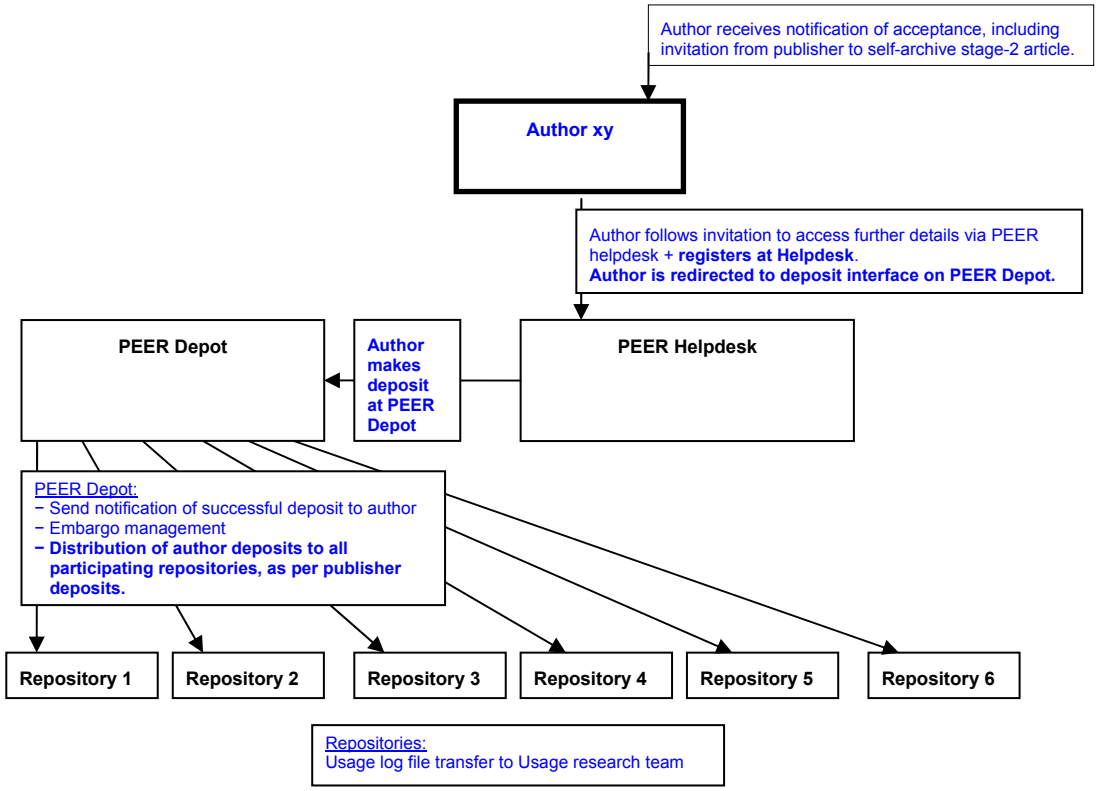

Figure 27: Scenario 3

Scenario 4: Central registration & central deposit a PEER Depot

- Author receives notification of acceptance, including invitation from publisher to self-archive stage-2 article. At this stage publication date is undetermined, and embargo period is unknown.
- Author follows invitation to access further details via PEER Helpdesk.
- *Author is redirected to the PEER Depot for registration and deposit.*

- *Author registers at the PEER Depot.*
- *Author deposits his/her data in the PEER Depot.*
- Author receives notification of successful deposit.
- Embargo management at the PEER Depot, as per publisher deposits.
- Distribution of author deposits to all participating repositories, as per publisher deposits.
- Repositories transfer their usage log files to the Usage research team.

Problems:

- Setting up a central author registration application means an additional effort of funding and staff resources.
- Central deposit seen as interference with 'natural way' of author deposit – referred to and approved by Executive.
- No communication between authors and repositories.

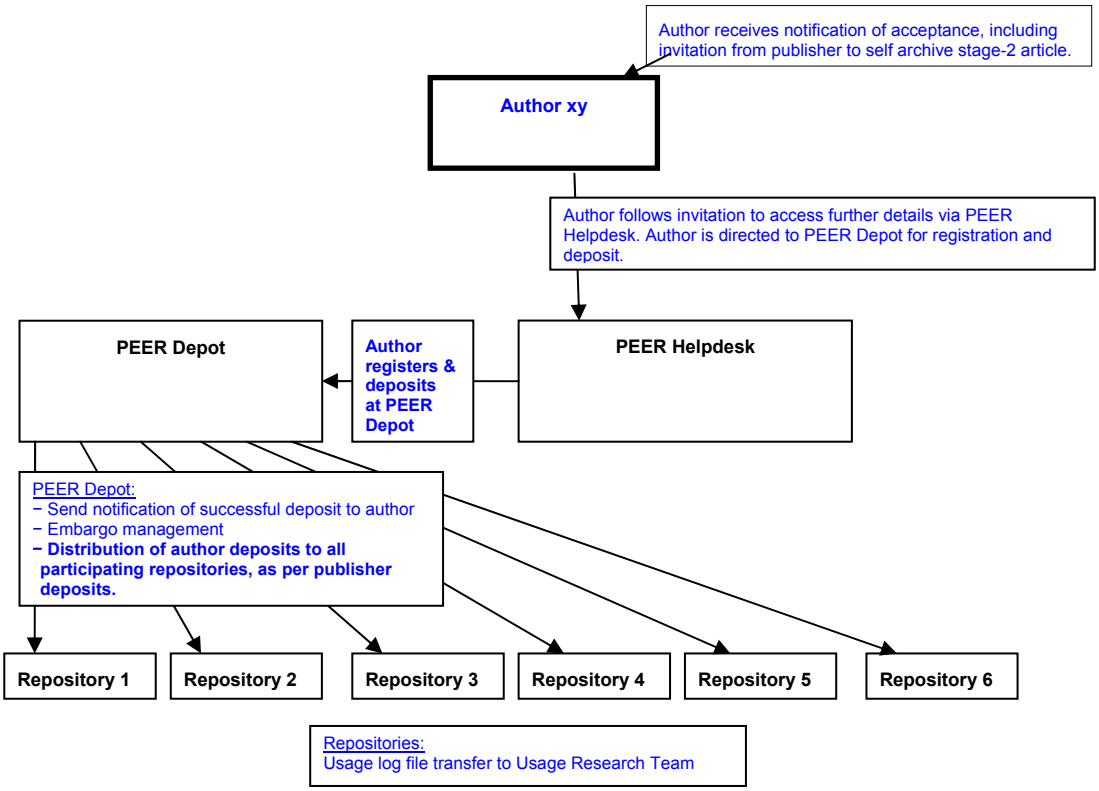

Figure 28: Scenario 4

Appendix F. Current and planned practice in the provision of usage data in a participating repository

PubMan@MPDL

PubMan@MPDL (http://pubman.mpdl.mpg.de/pubman/) is a participating repository in the PEER repository task force. As it is primary importance to limit additional effort imposed upon participating repositories, PubMan@MPDL is well positioned to provide a sample of current practices against which the PEER specification for the provision of usage data can be measured.

The PubMan@MPDL supports scientists and institutes in the management and the digital curation of their publications. This solution addresses all disciplines and focuses on the target groups of scientists, local librarians and local IT. It is built as an eSciDoc solution (see http://escidoc.org) that focuses on publication management. PubMan is used at present by several early-adopter Max-Planck Institutes. It is anticipated be used by all institutes within the Max-Planck Society (MPG) and to replace the current eDoc publication repository of MPG.

PubMan@MPDL provides the log files in the format NCSA combined.

Example:

134.76.162.XXX - - [01/Sep/2009:10:07:44 +0200] "GET /pubman/item/escidoc:265993:2/component/escidoc:265992/PEER_stage2_10.1080_slash_00268970903078542.pdf HTTP/1.1" 200 9193 "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.9.0.13) Gecko/2009073022 Firefox/3.0.13 (.NET CLR 3.5.30729)" "layout=PubManTheme; JSESSIONID=FC7627E3DED9E5F00E57C5861AC53A57"

The request contains the following information:

- Item identifier (escidoc:265993)
- File identifier (escidoc:265992)
- File name (PEER_stage2_10.1080_slash_00268970903078542.pdf)
- Session identifier (FC7627E3DED9E5F00E57C5861AC53A57)

Notes: Due to technical constraints, the file name has all slash symbols (coming from the DOI identifier) replaced with "_slash_" string. Due to privacy issues, the last three digits in the IP address of the request are be replaced with "XXX".

HAL@INRIA

HAL uses Apache and produces log files in the format NCSA combined.

BiPrints@UniBi

http://repositories.ub.uni-bielefeld.de/biprints/

BiPrints uses APACHE and produces log files in the format NCSA combined.

66.249.66.5 - - [12/Jan/2009:20:31:53 +0100] "GET /pdf_frontpage.php?source_opus=87&startfile=Egelhaaf_et_al_UniForsch2002.pdf HTTP/1.1" 302 414 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"