PEER D2.1 Draft report on the provision of usage data and manuscript deposit procedures for publishers and repository managers

Foudil Bretel, Christoph Bruch, Natasa Bulatovic, Wolfram Horstmann, Rianne Koning, Jacques Millet, Dale Peters, Maurice Vanderfeesten

ECP-2007-DILI-537003

PEER

D2.1 Draft report on the provision of usage data and manuscript deposit procedures for publishers and repository managers

Deliverable number	*D-2.1*
Dissemination level	*Public*
Delivery date	*31 March 2009*
Status	*Final*
Author(s)	*Foudil Brétel, Christoph Bruch, Natasa Bulatovic, Wolfram Horstmann, Rianne Koning, Jacques Millet, Dale Peters, Maurice Vanderfeesten*
Internal reviewer	*Gera Pronk, Jochen Schirrwagen, Julia Wallace*

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 report on the provision of usage data[1] and manuscript deposit procedures for publishers and repository managers sets out to establish a workflow for depositing stage-2 outputs in and harvesting logfiles from repositories to enable the research, conducted in work packages 4, 5, 6 & 7. It sets out an overall framework for including repositories, whereby a critical mass of content is made available in designated repositories to provide access to users.

The content comprises the contribution of approximately 11 publishers, who have agreed to participate in the project which aims to make available stage-2 outputs for 200 journal titles, in a research observatory. During the project more publishers will be invited to join the project to increase the number of journals to approximately 300 journal titles. To ensure that sufficient content is made available as a research sample to validate the research process, the publishers have agreed to collectively deposit 50% of the outputs on behalf of the authors. For the other 50%, publishers will invite the authors to self-archive their current manuscripts, and any previous manuscripts from participating journals.

1 Methodology

This report is the result of a process of negotiation with publishers to establish best practice in deposit procedures least disruptive of existing publication workflows, while minimizing interruption of repository ingest activities. As publication workflows differ considerably from one publishing house to another, the preceding negotiations were of great value in determining the minimum metadata requirements of the project, the point of intersection with individual publishers and the most effective mechanisms of transfer.

It was soon realized that the outcomes of the project could be best secured in the establishment of a closed intermediary repository that would receive the publisher deposit in the form of both 50% of the full-text outputs, as well as 100% of the metadata outputs, to serve as a base line control for the research process. The PEER Depot is conceptualized in a number of principles:

- The PEER project will establish and maintain a central depot to receive agreed deposits from participating publishers in specified formats.
- The PEER Depot is hosted by INRIA, and undertakes to ingest, normalize and transform the content data for distribution to designated repositories.
- The PEER Depot will manage the distribution according to precise delimitations of the PEER metadata set including the identification of European authors and specified embargo release dates. These mechanisms are detailed in Chapter 1.

The PEER Depot represents the defining role of the participating repositories in work package 2 and is designed to satisfy a number of requirements:

- serve as an intermediary between publishers and repositories, to minimise the manual work in data transfer
- maximise the number of repositories that can readily participate without significant effort outside of the project

1 The DoW originally names this task „Harvesting of logfiles“. Since it is not clear that “harvesting” will become the recommended practice, it is preferred in this document to call it “provision of usage data”.

- maximise the usage data they can provide
- minimise duplication of effort within the related project work packages 2 & 3

The methodology that circumscribes the author deposit was modeled repeatedly by members of the work package, in an attempt to secure the highest possible submission rate. This aspect remains an area of concern, as the erratic behavior of researchers in depositing personal outputs to repositories is an acknowledged threat to the success of the research process. A low author deposit rate could result in a sample too small to provide valid results. This issue is addressed in Chapter 2.

2 Repository Task Force

A repository task force comprising participating repositories will make available the content defined in the project. Four repositories are participating in PEER as partners and three additional repositories were identified by eIFL.net as potential members of the repository task force to extend the geographic coverage of the project. This has potential implications for the extension of usage evaluation criteria in the research activities of the project. In addition, the e-Depot at the Koninklijke Bibliotheek in The Netherlands was invited to act as a closed preservation repository, 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 content. The designation of a dedicated preservation agency is in accordance with DRIVER recommendations and through the University of Goettingen, links will be developed with the DRIVER project to utilise the guidelines and services enabled by the DRIVER infrastructure. The following DRIVER-compliant repositories have been invited to join the repository task force, and will participate in the development of guidelines for repositories foreseen in D3.1.

- PubMan, Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. (MPG)
- HAL, Institut National de Recherche en Informatique et en Automatique (INRIA)
- Göttingen State and University Library (UGOE)
- BiPrints, Universität Bielefeld (UNIBI)
- Kaunas University of Technology, Lithuania
- University Library of Debrecen, Hungary
- E-Depot, Koninklijke Bibliotheek, The Netherlands

3 Interaction between stakeholder groups

In the process of negotiation, issues relevant to the deposit procedures between publishers and repositories were tabled in two consecutive meetings, held on 7 November 2008 and 12 December 2008. These meetings reflect a consultative process conducted at two levels. Firstly, the joint meetings provided direct interaction between representatives of the PEER work packages 2 and 3, and representatives of the participating publishers. Secondly, the interval between the two meetings provided adequate opportunity to consult with their respective stakeholder group. The publisher group met separately following each meeting, and the work package members consulted with library and repository managers via the DRIVER consortium. This report therefore reflects a valid process towards a mutually acceptable level of consensus.

A number of issues could not be immediately resolved, where expected outcomes are uncertain, i.e. author response to an invitation to participate in the PEER project. The assignment of a persistent identifier (DOI) and work package inter-dependencies were flagged for ongoing consideration, and will be monitored and addressed in ongoing work of this work package.

4 Relationship between work packages and dependencies

Representatives of the PEER work packages 2 and 3 were invited to participate in the first meeting between stakeholder groups, to ensure continuity of the workflow from publishers to repositories, and the implementation of appropriate procedures with minimum duplication of effort.

In the course of the interaction, interdependency was identified between work packages 2 & 3 on the one hand, and work packages 1 & 5, on the other. To enable the research, a critical mass of content will be amassed in repositories to provide access to users. The manner in which this content is made available in repositories, and the nature of the logfiles specified to report on usage of that content is expected to impact on the methodology and outcomes of the usage research. This dependency is detailed in Chapter 3, and confirms the need to update this draft report in a succeeding deliverable, entitled D.2.2 Final report on the provision of usage data and manuscript deposit procedures for publishers and repository managers, which is due in Month 12.

5 Support mechanisms

This draft report sets out specifications for deposit procedures for both publishers and authors, and the reporting in logfiles of subsequent usage. Since this report is presented in draft format, it is anticipated that such specification will be adjusted as a result of actual implementation experience. Until the ultimate formulation of specifications and guidelines is achieved, a support mechanism is envisaged to assist both publisher and repository communities to share the experience gained, as detailed in Chapter 4.

This report is designed to be preliminary investigation of the issues addressed herein. As such, it forms the basis for a common understanding of the expected outcomes of the PEER project, and it highlights the issues of concern that need to be monitored and evaluated in the final report. Significantly, it indicates workflows, procedures and best practices that will be explored towards the establishment of best practice shared by publisher and library communities to ensure the future of scholarly communication.

1. Content deposits from publishers to repositories

1.1 Convention

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.

1.2 Proposed 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 on behalf of the publishers. The PEER Depot will be hosted at INRIA with the responsibility for facilitating publisher deposit, and dissemination to repositories. The content will also be retained in the PEER Depot, in case of processing or delivery errors. The PEER Depot will receive 100% metadata and 50% full-texts of the content. 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 closed archive (not accessible, nor searchable).

A diagram of the PEER workflow shows the expected parallel paths of publisher deposit and author deposit:

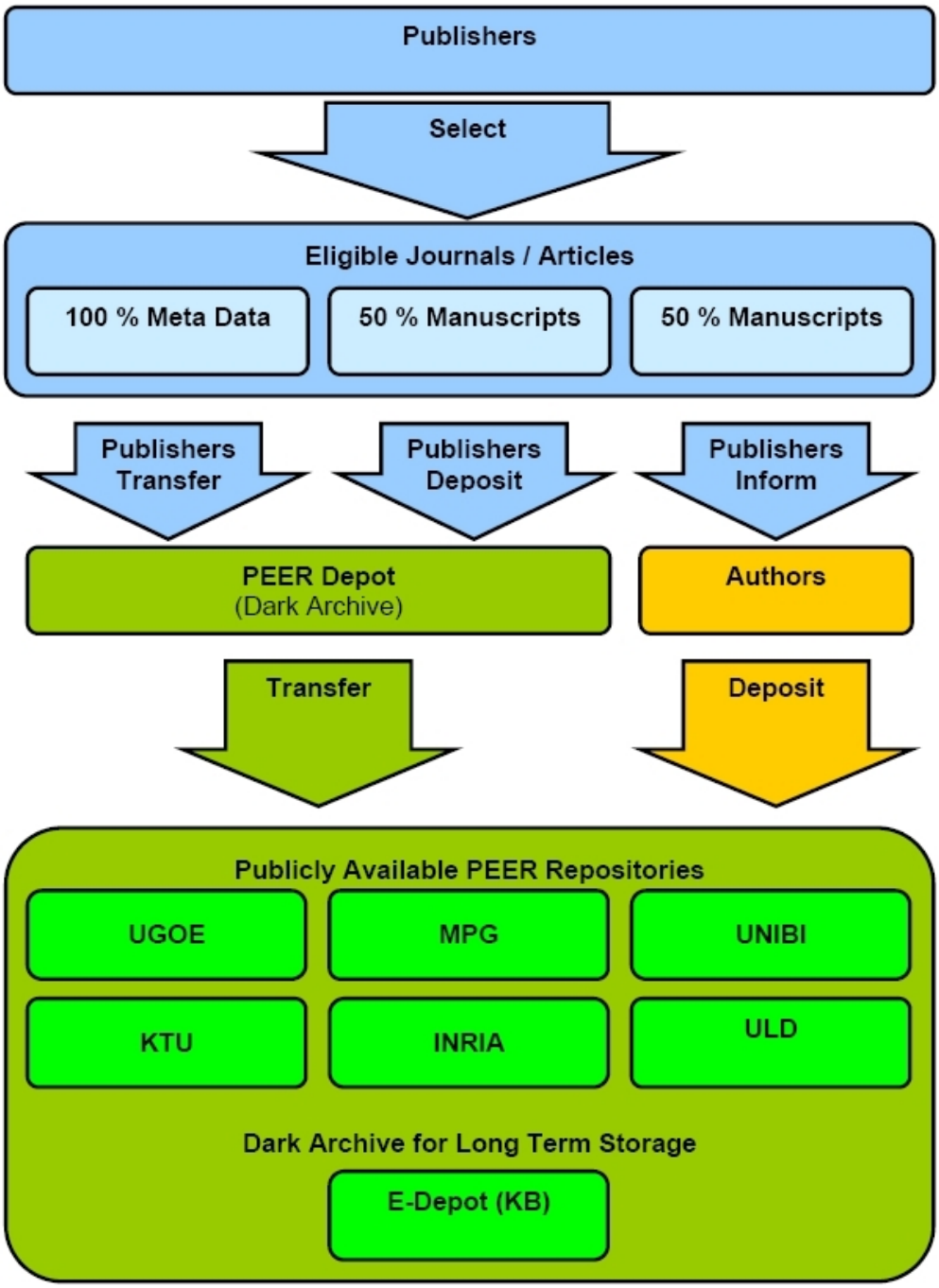

Figure 1: PEER workflow

1.3 Deposit procedures from publishers to the PEER Depot

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

- on a daily basis
- through FTP/S[1] into a dedicated directory
- as ZIP files, one per article

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

- 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.

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 acceptable for the PEER project. Very few publishers suggested they could only provide source files at stage-2, as LaTex or Word files. The conversion from source files to PDF would therefore take place at the PEER Depot. It would not, however, be possible for conversion from multiple Word files to a single PDF to take place at the PEER Depot.

Further investigation is therefore recommended into file formats that can be readily transferred from the publishers to the PEER Depot and further transferred from the PEER Depot to the repositories without undue additional effort. Suitable file formats would thus be supported by the PEER Depot (i.e. LaTeX/Word).

In order to identify articles, the full-text file received by the PEER Depot will be renamed as follows: "PEER_stage2_[full-formatted-DOI].pdf"[3] before submission by the PEER Depot to repositories.

1.3.2 Metadata

All Publishers agreed to provide metadata in an XML format. There seems to be a large acceptance of the NLM DTD [2] among Publishers. It is therefore best suited for PEER to conform to this *de facto* standard, and make use of the appropriate NLM tag set. A first analysis of the minimum metadata elements, set out below, showed that the Journal Publishing Tag Set [2] would be the most convenient. Further investigation of the use of the Journal Publishing Tag set for publisher deposit will be conducted in conjunction with the proposed use of the Text Encoding Initiative (TEI) for purposes of data management within the PEER Depot. *See Appendix A: TEI components of a PEER metadata format*

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 NLM DTD to TEIHDR (TEI Header). Where required, repositories can implement an ingestion tool based on the TEI. To avoid the transfer of this additional effort, the NLM-XML can also be passed to the repositories directly.

Since exports to the PEER Depot might occur in different systems at different stage in the publication workflow, Publishers indicated difficulties providing coherent stage-2 metadata. In some cases, critical metadata elements, such as embargo dates and a persistent identifier, 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. A system management convention will be defined for these three options, to determine a version control mechanism that indicates when a submission is complete.

---

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 2 digits, month, day, hour, minutes, seconds.

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

3 Full-formatted-DOI is the DOI (prefix and suffix), with the '/' character convert to '_'.

- 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*; a persistent identifier is provided in a second deposit during the embargo period; and matched against the publisher-article-id
- Option 3: the metadata updated by the publisher at stage-3 is submitted again, in replacement of the stage-2 deposit

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

Considering that the PEER project recommends the submission of as much metadata as possible, the minimum requirements are set out below.

DublinCore-like name	**Comment**
Title*	Article Title
Creator*	Corresponding Author's name: Last Name, First Name
AuthorEmail	Corresponding Author's email address
Description	Abstract
Date	Date of Publication
Identifier*	DOI or PublisherArticleId
Coverage	Geographic location of the Contributing Author: ISO 3166
Journal	Journal Title
Affiliation	multi-tier organisation list: Country, Organization, Laboratory
ISSN	These elements are not mandatory to electronic publication, and can be derived from CrossRef after DOI is provided, and may therefore not be provided by publishers. Further investigation will be conducted on the use of CrossRef for DOI resolution.
Volume	
Issue	
Page	
Type*	Default value = article. Mapped to *info:eu-repo/semantics/article, info:eu-repo/semantics/acceptedVersion*
Subject	Subject headings; Scientific classification (defaults to what is provided in the general STM Journal table)[1]

1 See 11.3 in the PEER Description of Work.

Language	ISO 639-3 (defaults to 'eng')
Embargo	Embargo Period (defaults to what is provided in the general STM Journal table)

Table 1: Minimum metadata requirements

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, and no further metadata, further investigation is required to source metadata from known public sources e.g. Public Library of Science (PLoS) or PubMedCentral.

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

1.3.3 Embargo period

The period of embargo determines the date of distribution from the PEER Depot to participating repositories. 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 = DeliveryDate

The PublicationDate is provided in the minimum metadata set, defined either at stage-2 or stage-3 deposit. The EmbargoPeriod, if not otherwise defined, defaults to that provided in the general STM Journal table[1].

1.3.4 Filtering

Two levels of filtering are envisaged as functions of the PEER Depot. Firstly, of journal titles for distribution to repositories, and secondly, of articles submitted by European authors. The PEER Depot is to receive 100% metadata and 50% full-texts. All selected content is to be disseminated to all repositories.

The selection of publisher-deposited full-text will be done at the journal title level, not manuscript level. The choice of eligible journal titles will be defined by the publisher community, with due cognizance of research requirements, viz. behavioural response of specific subject disciplines.

The project's design states further that only articles of European authors shall be included in the study. Since publishers do not generally filter content in this way, it was decided that the location of the corresponding author would be used to identify European content. The automated selection will take 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% of full-texts are to be disseminated to repositories, in fact, only that portion represented by the European corresponding author within that 50% are to be 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.

2 Deposit procedures from the PEER Depot to repositories

The most direct strategy for depositing into the repositories would be to implement a different ingest module for each repository. A preliminary study will be conducted to investigate the feasibility of automated ingestion, however, it is anticipated that each repository will provide its own ingestion service complicating this strategy.

Another approach, with long-term objectives, would be to unify the ingestion services, or agree on a common standard or practice, based either on the format used, or on the protocol used, like OAI-PMH or SWORD, as outlined below. The benefit would be to achieve a European core set of interoperable repositories capable in theory of accepting material directly from publishers beyond the project duration. Compliance with standard practice is achieved primarily through the DRIVER Guidelines, as implemented by repository managers.

2.1 Protocol Authenticated OAI-PMH

The PEER Depot could provide its content through OAI-PMH. This content would be harvested by Repositories. The PEER Depot would have to provide its own OAI-PMH format to include all metadata required by all repositories. In this respect, ***this solution is not extensible***, since each repository would imply a modification of the OAI-PMH PEER format. Since the PEER Depot is a closed depot, access would need to be restricted with HTTP Basic Authentication or SSL. Again, as for TEI, such effort is transferred to the repositories. It is therefore not recommended that additional mappings are made locally at a given repository.

2.2 Protocol SWORD

SWORD (Simple Web-Service Offering Repository Deposit)[1] is a JISC-funded lightweight protocol for depositing content from one location to another. It is a profile of the Atom Publishing Protocol and has the potential to become a standard deposit mechanism for digital content, and would facilitate push deposit from the PEER Depot to multiple repositories, with support for authentication over HTTPS.

Created to “lower the barrier to deposit”, a number of demonstration implementations have been developed, namely for DSpace, Fedora, EPrints and IntraLibary. The PEER Depot, as a potential SWORD-compliant deposit client, will investigate the implementation of the atompub profile, whereby repositories would have to implement the corresponding interface. The protocol can be implemented as an Application Programming Interface (API) to connect a service with a repository. A service can be a MS WORD add-in to deposit, a facebook widget, etc. For more information see http://www.swordapp.org/

Thus the SWORD deposit mechanism is extensible, it can be reused by repositories with legitimate potential additional effort for repository managers. In relation to the PEER workflow, the SWORD protocol could facilitate the deposit content from the PEER Depot into participating repositories.

The following diagram shows the SWORD interaction between the PEER Depot, the Repository system and the Repository manager.

1 http://www.swordapp.org/

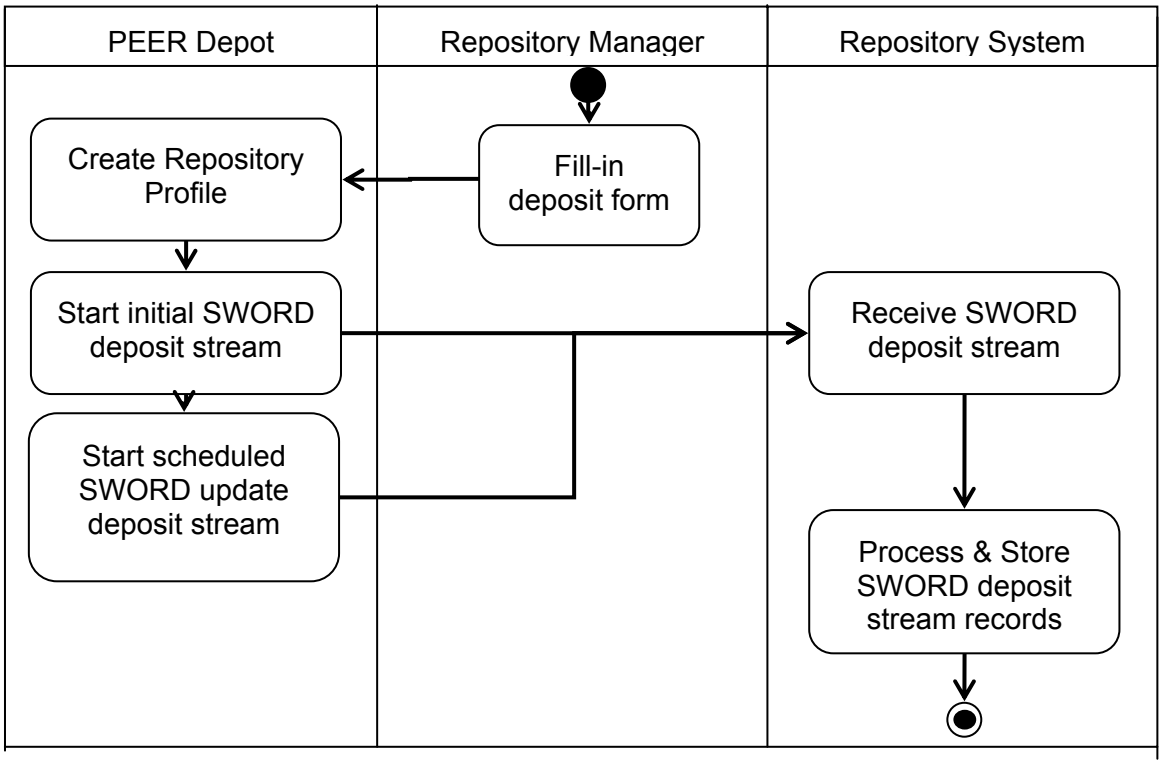

Figure 2: SWORD interaction between PEER Depot and repositories

The repository manager has to provide basic information in order to make deposit possible.

1. The URL of the SWORD receptor API of the Repository
2. The package format in which the repository system would like to receive the data. The standard packaging format is a ZIP-file with a METS manifest. Inside the Manifest the metadata format is SWAP. Additionally to the PEER project the metadata formats TEI and NLM can be used. The statements for the Packaging formats in the SWAP stream are required.
    a. http://purl.org/net/sword-types/mets/dspace (SWAP)
    b. http://purl.org/net/sword-types/mets/peer/tei (must be created)
    c. http://purl.org/net/sword-types/mets/peer/nlm (must be created)

The off-the-shelf repository systems Dspace, Fedora and Eprints have each a different standard of metadata to receive. Dspace would like to receive METS and SWAP metadata in a zip file. Fedora can handle any data stream since the data model can be customised, Eprints can be customised easily to handle any input.

The metadata formats that the PEER Depot should offer for the SWORD deposit procedure can be NLM XML (Journal publishing tag set), TEI metadata (for repository systems using this standard), and METS+SWAP.

For the repository the PEER Depot delivers each record of the second stage publication in a zip-file containing the metadata, the (converted) PDF-A and the complementary Stage-2 source files.

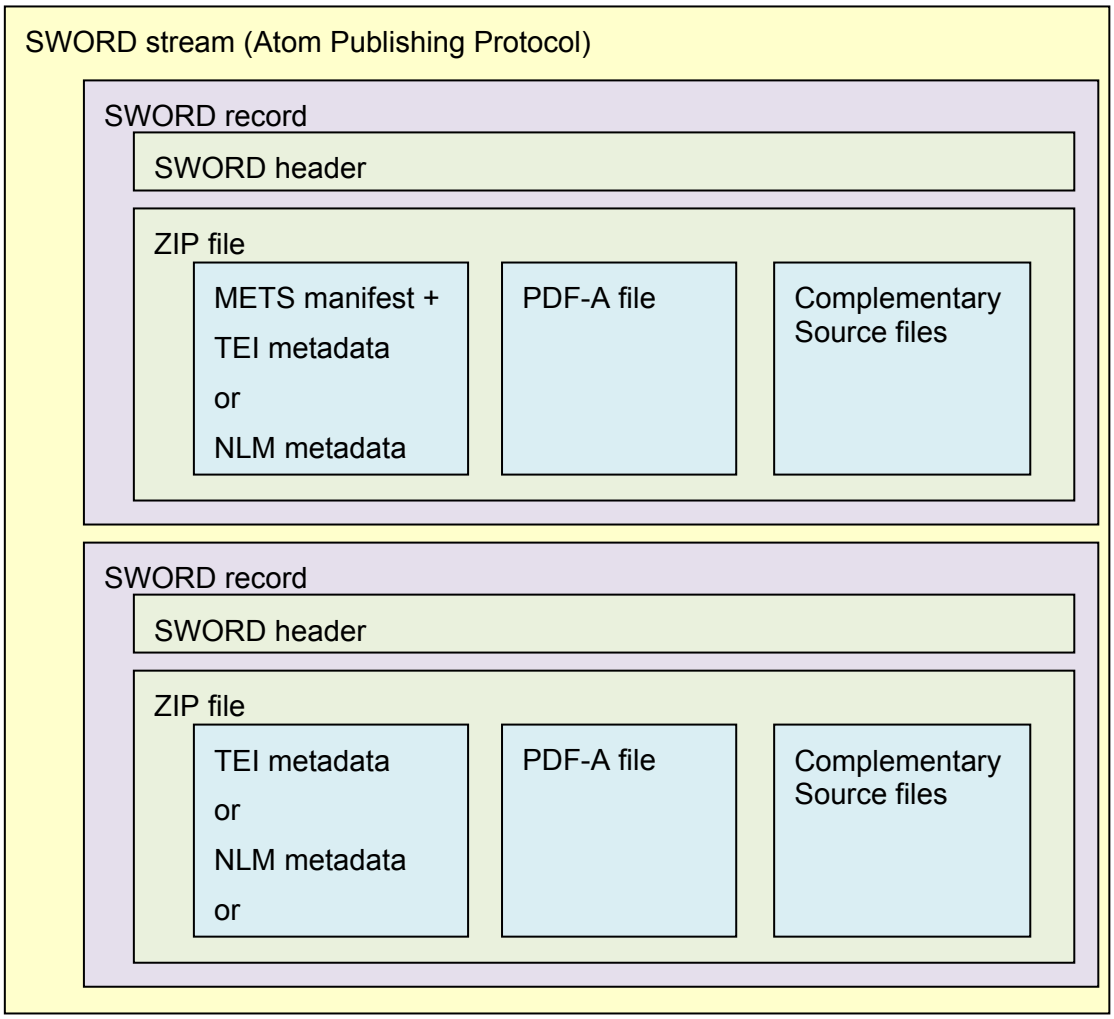

Figure 3: SWORD stream

Example of SWORD Atom feed expressed in XML:

```
<?xml version="1.0"?>

  <entry xmlns="http://www.w3.org/2005/Atom"
       xmlns:sword="http://purl.org/net/sword/">
    <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>
    <sword:userAgent>MyJavaClient/0.1 Restlet/2.0</sword:userAgent>
    <generator uri="http://www.myrepository.ac.uk/engine" version="1.0"/>

    <content type="application/zip"
       src="http://www.myrepository.ac.uk/geography-collection/deposit1.zip"/>
    <sword:packaging>http://purl.org/net/sword-types/mets/dspace</sword:packaging>
    <link rel="edit"
       href="http://www.myrepository.ac.uk/geography-collection/atom/my_deposit.atom" />
  </entry>
```

One can see the entries are delivered in chunks. This allows the receiving end to start downloading the zip-file when the information of each entry has reached the repository while the SWORD stream keeps coming in.

3 Deposit procedures from the PEER Depot to LTP Depot

3.1 Introduction

The e-Depot of the National Library of The Netherlands (KB) is aimed 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 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 for the manuscripts made available through the participating PEER repositories. 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 LTP Depot in which the data objects and the accompanying metadata are kept safe beyond the duration of the project. In line with KB's current policy, KB also renders on site access to stage-2 manuscripts via its KB catalogue.

See Appendix B: Technical specifications for LTP Depot

3.2 Content

As the e-Depot is based on long-term preservation of content objects, KB will only process those manuscripts deposited from the PEER Depot. The PEER Depot will receive 100% metadata and 50% full-texts. This means that PDD can send LTP Depot 50% full-texts including 50% corresponding metadata. LTP Depot cannot process and ingest the balance of 50% of metadata held in the PEER Depot. Authors will deliver max. 50% of the corresponding full-text directly to the participating repositories and not to the LTP Depot.

3.3 Workflow for Transfer to LTP Depot

The function of the LTP Depot is to preserve stage-2 manuscripts as deposited in the PEER Depot. Access is not the main objective and will be provided on site for KB pass holders only. 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.

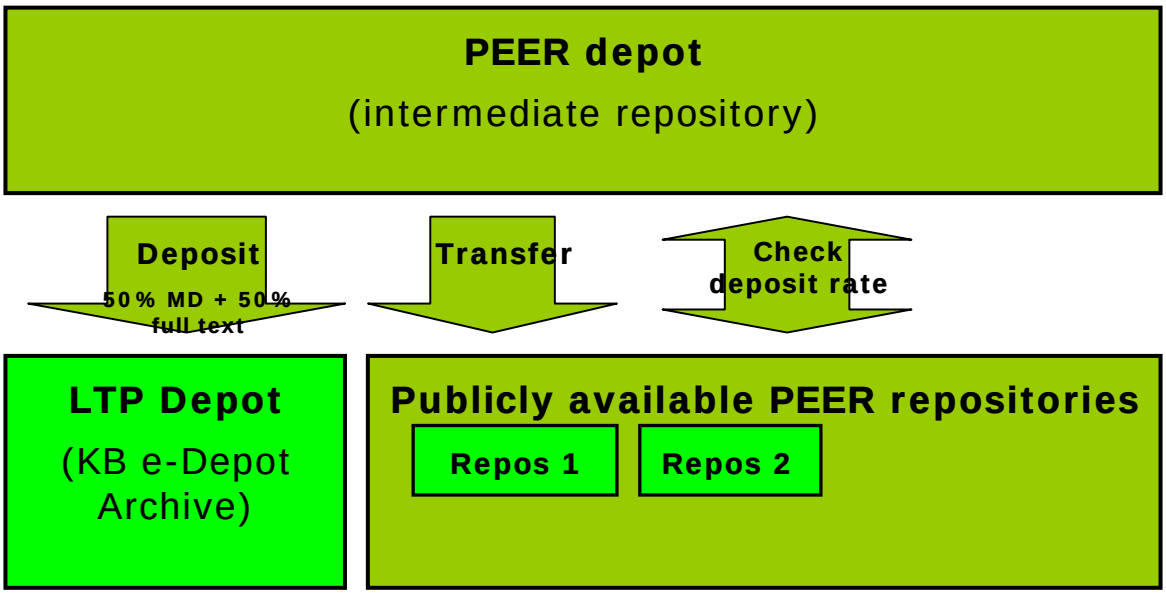

Figure 4: Workflow for transfer to LPT depot

As described in the workflow, stage-2 manuscripts from publishers will be transferred from the PEER Depot to LTP Depot. Before transfer takes place, the content of each zip file is converted[1] to its final stage for archiving. Each zip file contains:

- a main file in PDF
- the corresponding metadata file
- and possibly supplemental files with the article

The zip file will be transferred via FTP to the LTP Depot. In processing the content, bibliographic metadata described according to the PEER DTD will be converted to KB's DTD, whereas the original metadata delivered by the PEER Depot is stored with the content and converted metadata. The PEER data depot and the KB LTP Depot will agree on dedicated ingest processes, on delivery frequency and form of delivery. Processing of the packages is based on the OAIS reference model.

See Appendix C: KB use of OAIS Model

3.4 Metadata

Standardised metadata will enable the PEER Depot be easier to function as a catalogue. Another advantage of making use of one standard for the whole PEER workflow is the communication between the different depots; this will be more simplified. To achieve standardised metadata PEER will create its own specific DTD. Currently there are 2 options: using a Tag Set of the NLM DTD or the TEI DTD. As most publishers participating in the PEER already use the NLM-DTD, this is probably the best option to use as the standard PEER DTD.

See Appendix D: Analysis of the NLM DTD

---

1 Content is converted to the PEER DTD and text main files (the article itself) are converted to PDF (/A).

Publishers deliver the stage-2 manuscripts including related DTD to the PEER Depot. The Depot will convert the bibliographic metadata into the mandatory, recommended and optional elements of the PEER DTD.

See Appendix E: PEER DTD.

Critical to the LTP workflow is the interval where publishers may not be able to deliver complete metadata at stage-2, e.g. where a DOI or ISSN is missing. In this case, missing metadata elements would be sent to the PEER Depot in an additional XML file, at a later date. Identified as Option 2 outlined in 1.3.2 above, this procedure will entail the compilation of multiple XML files, together with the content (PDF file and possible supplementary files) in a single zip file for each article. Only when this final version is complete, is the archival submission package ready for transfer to the KB LTP Depot.

Furthermore, under Option 2, publishers have indicated that metadata may be delivered sequentially, according to the table below. The Acceptance state represents all metadata delivered in the first submission. Metadata described as Update will be delivered as an extra metadata file. Both files need to be combined in the final metadata file for archiving of the stage-2 manuscript.[1]

Acceptance (XML file1)	**Update (XML file2)**
Article title	Publication Date
Author	DOI
Author e-mail	Volume
Country/coverage	Issue
Journal title	Description /Abstract
Affiliation	Page
Publisher ID	Publisher ID
Subject	Embargo
Language (default=English)	
Type (default = article)	
ISSN (electronic and/or print?)	

Table 2: Sequential metadata transfer

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 this standards becomes more widely accepted.

1 PEER Technical Meeting 12-12-2008.

3.5 Digital Preservation

The PDI (Preservation Description Information) is required for adequate preservation of the Content Information. Besides bibliographic metadata, KB needs to identify metadata categories as specified under the OAIS model, listed below. For each category, KB prefers a separate format[1]. Currently the following categories and related 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.

The feasibility of the provision of OAIS–compliant metadata needs to be further examined. Besides the minimum metadata requirements set out in 1.3.4 and 1.5.4 above, the PEER project may well determine the need for more detailed metadata, for example according to IR workflows, long-term preservation aspects, etc. To facilitate this examination, all noted metadata elements for the PEER DTD (for both publishers and KB LTP Depot) are mapped to the NLM elements.

See Appendix F: KB DTD

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

1 OAIS defines for these 4 subcategories which are all types of metadata.

2 Content deposits from authors to repositories

Authors eligible for participating in the PEER project will be notified via the publisher cooperating with PEER. While the author response is unpredictable, this work package sets out to facilitate the process of author deposit, as far as possible, without interfering in established practice. Authors will therefore be encouraged to follow their established practice of deposit in an institutional or subject-specific repository. Failing such practice, deposit in one or more of the PEER designated repositories will be recommended. It should be noted that is highly unlikely that authors would be willing to deposit twice. An author deposit procedure parallel to that of publisher deposit is not possible, without undue intervention in scholarly practice. Precisely this lack of a controlled author deposit procedure will determine 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 deposit the stage-2 manuscript, either per established practice, and/or in one of the participating PEER repositories. The request for deposition will include a request to inform the project, should the author intend to deposit the manuscript in a repository other than one of the specified PEER repositories.

2.2 Communication with authors

For reasons of data privacy, the participating publishers are not able to make available the contact details of eligible authors, and no direct communication is envisaged. However, since it is expected that authors may choose to respond immediately upon receipt of invitation to deposit, the invitation will be linked to a dedicated page on the PEER website. An online interface will be established to guide authors through a simple deposit procedure; to outline rights and embargo issues; to capture the URL of the repository used for author deposit, and redirect to participating PEER repositories, where required. The lack of direct communication unfortunately precludes any reaction to failed author submission, whether by oversight or by technical error.

2.3 Monitoring author response

Deposition will be monitored by the behavioural research undertaken in WP4, and measured against the 100% metadata control managed by the PEER Depot.

While it is not possible to predict the behaviour of authors invited to deposit, the following diagram attempts to illustrate the anticipated author deposit workflow. 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.

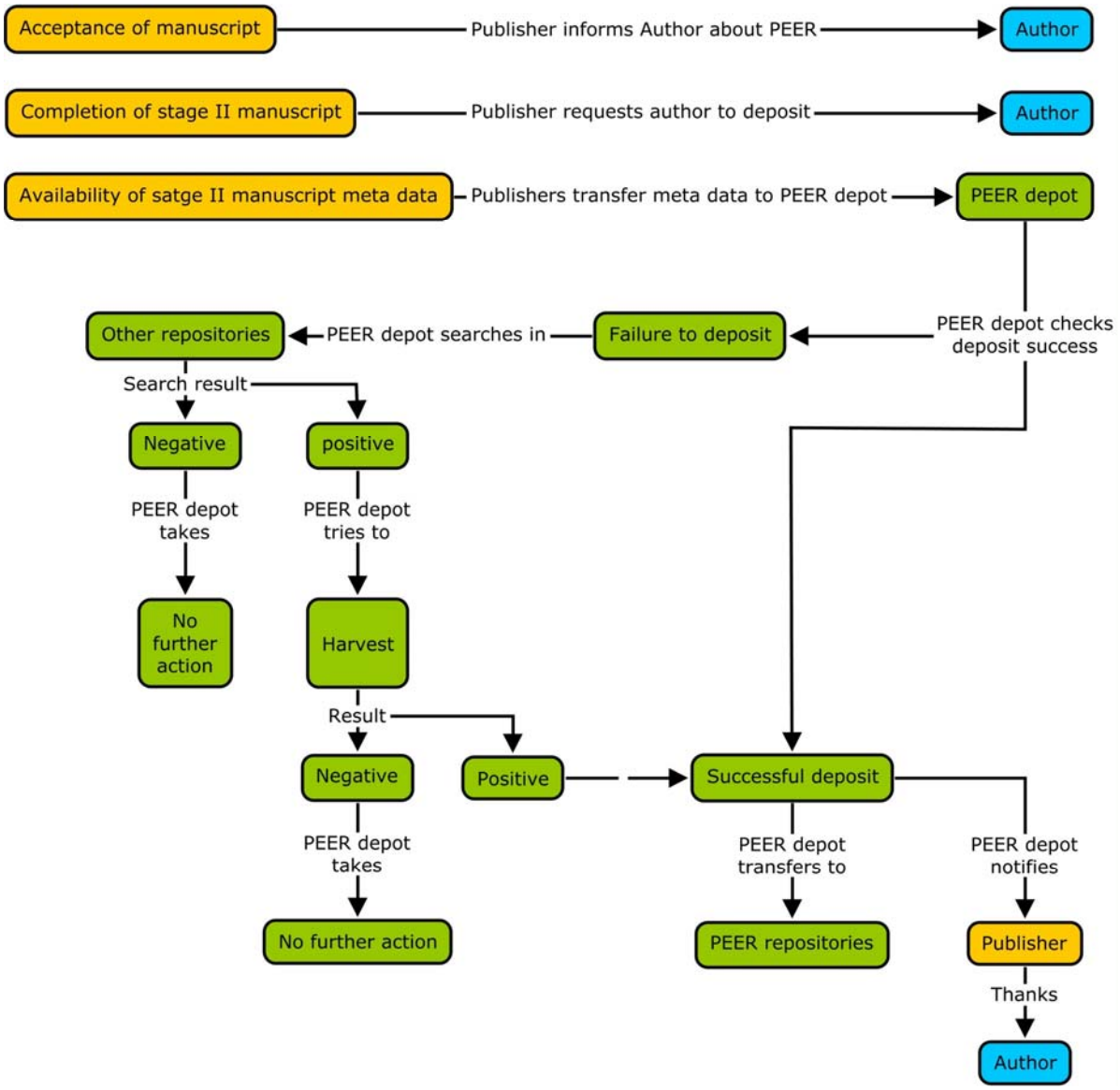

Figure 5: Author deposit workflow

3 Provision[1] of usage data

3.1 Introduction

This chapter defines the usage data that providers (i.e. publishers[2] and repositories) should provide to enable research on usage statistics. According to decisions taken in the PEER project a very basic solution is presented: PEER repositories participating as usage data providers should monthly upload to an FTP-Server provided by PEER an ASCII-file containing all usage events complying with the format commonly elaborated as "NCSA combined"[3], where the filename of the PDF provided by PEER, defined by the specification of PEER-deposit-procedures (see chapter 1), is represented in the http-request.

Those parties selected to perform usage research in WP5 will be required to approach publishers individually for access to their logfiles. This interaction with publishers will not be described further in this report.

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 will be launched for behavioural and usage (December 2008) as well as for economic research (Summer 2009), respectively. In order to enable this research organized in Work Package 1 (WP1) 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, and

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, which is described in chapter 2, and
- recorded usage events for these publications from all participating repositories as data-providers. The remainder of this document describes how these requirements can be met by the PEER project, specifically WP2 and WP3.

3.1.2 Status of this specification

This is a draft version towards the fulfilment of D2.1 in January 2009. Further elaboration of the specification in 2009 might be subject to the results of the research tender, more specifically to the parties selected to perform usage research.

---

1 The DoW orginally names this task „Harvesting of logfiles". Since it is not clear that "harvesting" will become the recommended practice, it is preferred in this documents to call it "provision".

2 The proposed procedure is not yet aligned with publishers since it was originally assumed that this would be organized by STM. It became clear recently that both repositories and publishers should use the same routines for usage data provision. Thus, the alignment process with publishers is still to be undertaken

3 Logfile formats:
http://publib.boulder.ibm.com/tividd/td/ITWSA/ITWSA_info45/en_US/HTML/guide/c-logs.html#ncsa

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 too much 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 logfiles to the party acquiring the usage research tender. Thus, specific questions to be answered by this document are limited as:

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

3.2 Transmission of Logfiles

It is proposed here that local data providers transmit their local logfiles monthly to a central facility that is provided by PEER. The research performing party shall then receive access rights to this central PEER facility. Since a similar transmission procedure is planned to be applied for publications by publishers (“PEER Depot”) and this one will be based on the FTP protocol, it is proposed to apply the same principle for usage data. Thus, local data providers will receive instructions to send their logfiles to a specific FTP-server (supporting FTP/S). The reader may picture this package as a ZIP-file with the naming convention: "*PEER_usage_[data_provider_name]_[yymmddhhmmss].log*" to avoid mistaken file overwriting.

3.2.1 Structure of Logfiles

It is generally proposed to deliver logfiles as raw and as comprehensive as possible [7]. This implies to refrain from applying cleaning routines like those applied in analytic tools such as AW-Stats [8]. Also, it is not assumed (according to [6,7]) that logfiles may contain non-PEER documents, since filtering out PEER documents is an obligation of the research performing party (see also “Identification of Documents”). [11].

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

attribute	mandatory /optional	example	comment
host	m	125.125.125.125	maybe anonymized[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	o	200	
Bytes	o	1043	
referer	m	http://www.google.com/	Highly recommended
user_agent	o	"Mozilla/5.0“	

Optional fields that are missing must be represented as “-“. Logfiles 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)"
```

As an apache configuration:

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

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 logfiles and some pitfalls are manifest:

- The filename or identifier appearing as http-request in the logfile may only be known to the application (repository) but has no reference to PEER documents.
- In other cases raw logfiles may contain only cryptic calls of services (e.g. a PHP script[3] Web-services, Session Management, Cookie etc.) that does not contain any identifier and render a later identification of documents difficult or impossible.

---

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 logfile has to be made anonymous before submitting them (procedures and software to be specified).

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

3 192.168.47.11 - - [15/Jan/2009:07:35:06 +0100] "GET /sendfile.php?type=0&file_id=8c49d37b913076c63054db5414d545c0 HTTP/1.1" 200 61846

- When http-'post' is used instead of http-'get' the identifier may be used as 'TYPE=HIDDEN' and do not appear in the logfile.

These cases – let alone the many others that can occur – would render it impossible for the research performing party to infer which usage event belongs to a specific PEER-document. Thus, it will be required to allow for specific elaborations of the logfiles that are to be prepared by an individual data provider. These elaborations might have different formats and encodings (e.g. TXT, CSV, XML, XLS) but it is highly recommended to use simple Ascii-textfiles in order to avoid errors in the post-processing by the research performing party and keep workload for them minimal.

This minimum recommendation may be refined after all the initially participating repositories have provided their current policies in logging usage events and after those results are analysed.

3.3 Identification of documents

Raw logfiles will contain much data of no relevance to the PEER project. Even though it has been decided by WP1 to leave the task of filtering out that data that are relevant for PEER to the usage research performing party, it is assumed responsibility of WP2 to indicate the identification of the usage 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 logfiles*.

As decided in the PEER executive committee [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 for the usage research performing party (e.g. through an OAI-PMH publisher), in order to obtain the current list at any given point in time, enabling the matching between usage events in logfiles and eligible articles.

It has also been agreed (for other reasons that are not explained here [12]) that an identifier will be created at the PEER Depot that reflects the publisher, the journal, and the article, e.g. in the form:

[publisher]_[journal]_[publisher_internal_article_identifier].

This *filename* of the full-text provided by PEER should, in an optimal situation, allow easy track usage events in the logfiles. It is therefore mandatory for participating repositories to represent this PEER-filename, either in the URL of the document, or as an additional filename in the metadata.

However, since it was also decided [6;7] to have only 50% of the articles eligible in PEER deposited on behalf of the publishers while the other 50% are subject to spontaneous author submission, the latter 50% will be not easily identified: the usage research performing party could have to match metadata collected from repositories with the list of articles eligible in PEER for building correspondent pairs and find a way of identifying how usage events related to the spontaneous deposited articles are represented in the raw logfiles. The alternative suggestion, of using 100% of the full-text, tagging it with an identifier that can be tracked in usage data and offering this for author deposit was declined by the project executive committee [6;7], on the grounds that the aims of the project would be undermined if this proposal was followed. This report thereby notes the potential risk of performing usage research only on the 50% that are publisher-mediated deposits.

It is recommended however, that the deposit procedures within the participating repositories ensure storing PEER identifiers of as additional identifier. Thus as a minimum, a list with pairs of local identifiers and PEER identifiers can be separately provided by the participating repositories, to help researcher team identifying relevant data from the logfiles.

3.4 Expected Result

The expected result of this procedure is a single-point service (such as an FTP/S server) provided by PEER, by which the usage research performing party can collect all raw server logfiles that contain usage events of PEER articles from participating data providers. 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] Pre-announcement of call for tenders for research sent by Chris Armbruster

[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

[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

The nature of this draft report requires some consideration of the means of ongoing communication between the publisher community, the PEER Depot, and the repository community to resolve the issues raised towards a final report scheduled as D2.2 in M12.

An agreed point of communication between members of the work package has been established in the WP2/3 listserv at: peer-wp2-3@inria.fr and the Project Manager, who serves to represent the publisher community, is included in the listserv communication mechanism.

As the draft recommendations of this report are tested in the course of the project, the queries that arise in areas of concern indicated, as well as other related queries will be documented towards the formulation of the final report.

4.2 Establishment of a helpdesk

Actors involved in the ongoing support facility envisaged include Authors, Publishers, Repository representatives and PEER researchers. All these actors will be likely to call upon a support facility. Therefore, a telephone hotline would overly burden the support team. As indicated in Chapter 2 above, an online interface will be established to guide authors through a simple deposit procedure.

This online interface will be established as a central point of author support on the PEER website, and maintained by a team formed of the technical representatives of WP2/3. This support team will rely on, and possibly forward requests to designated representatives in each community.

The helpdesk will also provide support for repository managers, with for example, information for how to obtain the “NSCA combined” logfile 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.

Technically, the support facility may be implemented in the form of a ticket system (such as *Trac* or *Request Tracker/RT*). In addition, frequently asked questions (FAQ’s) will be developed and published on the helpdesk site, based on the successful precedent set in DRIVER, at: http://helpdesk.driver.research-infrastructures.eu/

A ticketing system is highly effective since the questions and answers are well documented. The results can be published, and the participants are able to review issues that arise. Where the ticketing system is made public, the “wisdom of crowds” principle can be applied to gain more efficient response to complex problems.

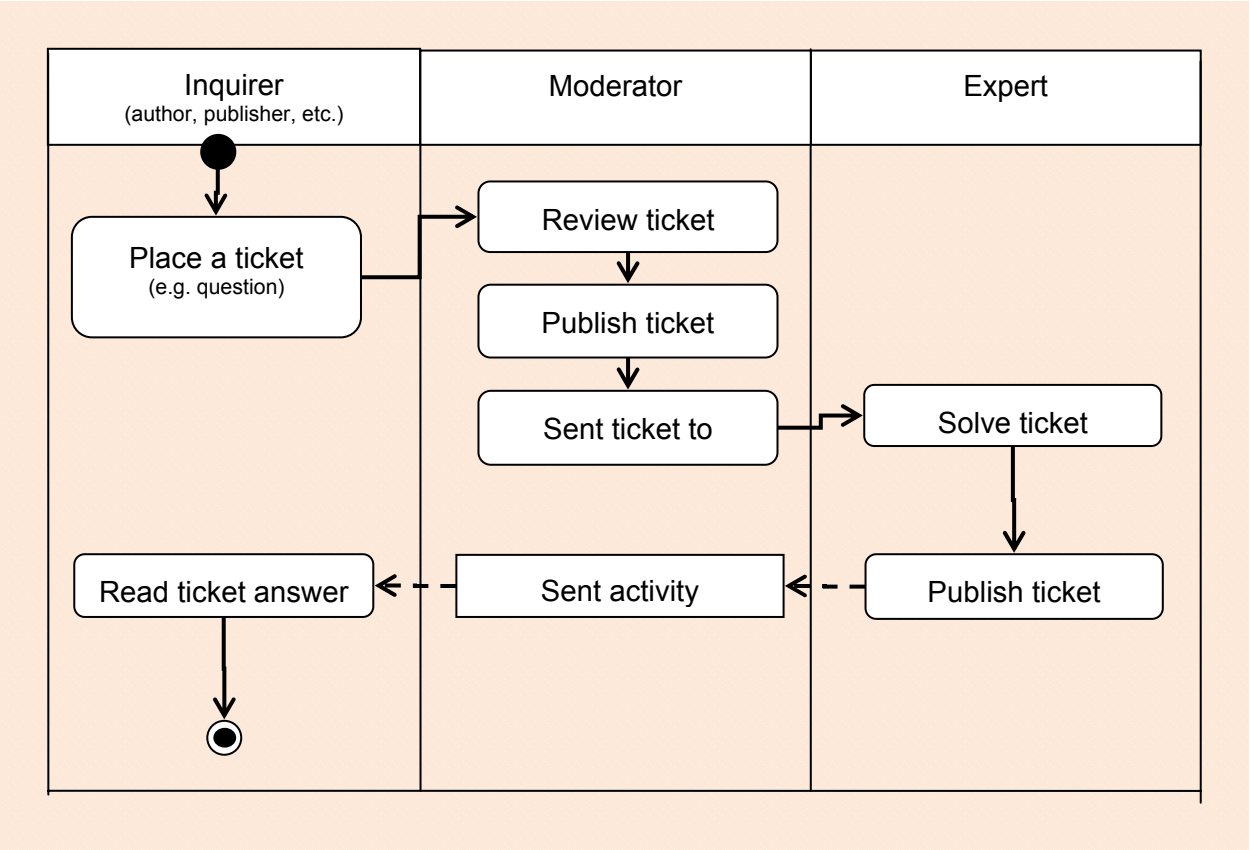

Figure 6: UML activity diagram of helpdesk ticketing system workflow

5 Conclusions

The investigation undertaken towards this draft report, D2.1 is the outcome of a valid process of consultation between representatives of the work package, and representatives of the participating publishers. As such, the report reflects a team-based approach to the investigation that is appropriate to the PEER project, which brings together for the first time, disparate interest groups from the publisher, library and repository and the research communities. While in effect, a number of technical issues that were raised for discussion remain unresolved, these were noted for monitoring and further consideration in the final report, D.2.2. Good co-operation was thus established during this process that will facilitate the ongoing interaction between publishes and repositories in the course of the project.

It should be noted that the task did not require consultation with the research community, acting in PEER as representatives of the authors. The dependency of this work package on the envisaged research processes was also noted. Interaction with research tenders was not possible at the time of writing, and as a result, the author deposit workflow remains conjectural, and the need has been expressed in Chapter 2, on author deposit, to interact with parties selected to perform behavioural research in WP4 and in Chapter 3 on the provision of usage data, with parties selected to perform usage research in WP5. It is expected that specific areas of this draft report will require amendment as a result of such interaction, for documentation in a final report in D2.2.

In addition, a number of specifications set out in this report, such as the assignment of a persistent identifier (DOI), and provision of OAIS-compliant metadata needs to be further examined. These issues will be tested in the interim and amended according.

Furthermore, some questions relevant for this work package are outside the scope of this chapter, but noted here as a result of discussion.

- Parties selected to perform usage research in WP5 will be required to approach publishers individually for access to their logfiles.
- Exchange agreement will be required between PEER and external repositories: “get publication data for giving usage data”?
- Confidentiality agreements will be required, (a) between a participating repository and PEER and (b) between PEER and the research performing party.

Finally, the effort represented by partners in this work package is duly reflected in the report as a comprehensive response to the project objectives to investigate the effects of the large-scale deposit of publications in repositories on user access, author visibility, and journal viability. While this report sets out technical issues of data transfer between publishers and repositories and between authors and repositories, it is not limited in scope to technical issues, but also reflects every best effort towards the overall success of the project, and an exercise in building good will between PEER stakeholder communities.

Appendix A. TEI components of a PEER metadata format

This annex describes the output format that will be adopted by the PEER transient repository (Partner: INRIA) for distributing the metadata information provided by publishers. It is based on the TEI[1] guidelines, with some additional constraints intended to make the corresponding information structures universally interpretable.

Note: this part will be revised when the actual implementation requires an expansion of the encoding scheme to deal with extra-feature.

1 Overview

The proposed structure combines a global structure (<TEI>[2]), which can potentially integrate any information that can be found in a full-text representation of a paper article, and a sub-structure (<biblStruct>[3]) that specifically contains the bibliographical information of the article. This allows us to process in a uniform way the two following scenarios:

1. We receive full paper articles (or retrieve them from repositories such as PMC) and convert them to the TEI format, thus exploiting all its expressive capacities;
2. We receive specific metadata information, with possibly some additional content (e.g. abstract). We then create a highly simplified <TEI> structure, which is mainly a container for disseminating the bibliographical content.

In the remaining part of this document, we will primarily address the second scenario, which is the one needed for the research to be carried out within the PEER project.

2 Representation of bibliographical information

The representation is based on the TEI <biblStruct> element which is organised as follows:

```
<biblStruct type="article">
        <analytic>
                …
        </analytic>
        <monogr>
                …
                <imprint>
                        …
                </imprint>
        </monogr>
        …
</biblStruct>
```

A <biblStruct> is mainly divided into two sub-structures:

- <analytic> to indicate the bibliographical characteristics of an article (title and authors);
- <monogr> to account for the publication details of the journal (journal name, publisher information, issn, etc.), and contains in turn a <imprint> element which

1 Text Encoding Initiative (www.tei-c.org)
2 http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-TEI.html
3 http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-biblStruct.html

gathers publication and/or distribution aspects of the article in the corresponding journal (pagination, volume, issue, etc.);

- When applicable, additional notes or identifiers can follow, for instance, the DOI, pubmed central id or repository specific id will appear here:

```
<biblStruct type="article">
        <analytic>…</analytic>
        <monogr>…</monogr>
        <idno type="pmid">12345678</idno>
</biblStruct>
```

3 The <analytic> element

a. Overview

The title of a journal article is represented by means of the <title> element (with appropriate @level attribute) as follows:

```
<title level="a">Multilocus Analysis of Age Related Macular Degeneration</title>
```

When necessary a further @type attribute may be used to differentiate between main and subtitles (@type=”main” vs. @type=”subordinate”).

Each author in the <analytic> element is independently described by means of an <author> element. This element contains the author’s name, affiliation and addresses − when available − as presented in the outline below:

```
<author>
<idno type="...">...</idno>
        <persName>
                <forename>Michael</forename>
                <surname>Dean</surname>
        </persName>
        <affiliation>…</affiliation>
        <email>dean@ncifcrf.gov</email>
</author>
```

b. Dealing with affiliations

The <affiliation> component of <author> is intended to contain any potentially relevant information with regard to the author’s academic situation: research group, laboratory, institution.

```
<affiliation>
        <orgName type="laboratory">CSA Department</orgName>
        <orgName type="institution">Indian Institute of Science</orgName>
        <address>
                <settlement>Bangalore</settlement>
                <postCode>560012</postCode>
                <country>India</country>
                <addrLine type="phone">+91-80-22932386</addrLine>
                <addrLine type="fax">+91-80-23602911</addrLine>
        </address>
</affiliation>
<email>kavitha@csa.iisc.ernet.in</email>
```

4 The <monogr> element

The <monogr> element gathers journal identification information (journal title and ISSN together with the publishing information contained in its <imprint> sub-element). For instance:

```
<monogr>
        <title level="j" type="main">European Journal of Human Genetics</title>
        <title level="j" type="nlm-ta">Eur J Hum Genet</title>
        <idno type="ISSN">1018-4813</idno>
        <imprint>…</imprint>
</monogr>
```

5 The <imprint> element

By imprint is meant all the information relating to the publication of a work: the person or organization by whose authority and in whose name a bibliographic entity such as a book is made public or distributed (whether a commercial publisher or some other organization), the place of publication, and a date. It may also include a full address for the publisher or organization. Full bibliographic references usually specify either the number of pages in a print publication (or equivalent information for non-print materials), or the specific location of the material being cited within its containing publication.

The <imprint> element is organised as follows:

```
<imprint>
        <pubPlace>Oxford</pubPlace>
        <publisher>Clarendon Press</publisher>
        <date>1969</date>
        <biblScope type="vol">3</biblScope>
        <biblScope type="issue">2</biblScope>
</imprint>
```

The possible values for the attribute @type on <biblScope> are the following:

- vol: volume
- issue: issue
- fpage: first page
- lpage: last page
- pp: number of page when the information about full pagination is not available[1]

6 Full <biblStruct> skeleton

```
<biblStruct type="article">
        <analytic>
                <title level="a">…</title>
                <author>
                 <persName>
                     <forename>…<forname>
                     <surname>…</surname>
                 </persName>
```

1 We restrict here the semantic of the recommended value. (cf. http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-biblScope.html)

```
                    <affiliation>
                       <orgName type="">…</orgName>
                       <address>…</address>
                    </affiliation>
                    <email>…</email>
                </author>
        </analytic>
        <monogr>
                <title level="j">…</title>
                <idno type="ISSN">…</idno>
                <imprint>
                 <publisher>…</publisher>
                    <pubPlace>…</pubPlace>
                    <date>…</date>
                    <biblScope type="fpage">…</biblScope>
                </imprint>
        </monogr>
        <idno type="DOI">…</idno>
</biblStruct>
```

References

http://www.idealliance.org/papers/extreme/proceedings/html/2002/Rosenblum01/EML2002Rosenblum01.html

Appendix B. Technical specifications for LTP Depot

To ensure that content used for research in PEER will remain accessible in the future, stage-2 manuscripts will be archived for long-term preservation in LTP Depot. PDD will deliver publishers' content to the LTP Depot. KB handles guidelines regarding the following subjects:

Batch structure

Batches should be delivered in .zip format. KB prefers to receive all content of one article per zipfile without any other directories or compressed or decompressed files in it.

Name convention files

All files which belong to one article need to have a unique and identical name convention. For example: one article exists out of 1 PDF, 2 Word supplements and 1 XML file. The following file name convention can be handled:

publisherx-10.1177/152216280100400120-20080912101020.zip

The following alternativefile name convention is probably too long:

[PublisherArticleID].[yymmddhhmmss].zip

Within the above mentioned zipfile there are 4 files:

publisherx-doi-20080912101020.pdf

publisherx-doi-20080912101020-s1.doc

publisherx-doi-20080912101020-s2.doc

publisherx-doi-20080912101020.xml

Besides the above described example, KB prefers not to handle: underscores, interspaces, slashes, backslashes, commas or dots in a name convention, but a hyphen.

PDF Guidelines

The main file of the article itself needs to be a PDF. Regarding PDF files, KB handles guidelines which are mainly about the following subjects: Accessibility and structure, Fonts, Compression, Images, Executable actions and Colour. 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. For more detailed information read Appendix C: *KB use of OAIS Model*.

Metadata

Metadata needs to be in XML format including Unicode UTF-8 character encoding. Mandatory, recommended and optional elements are described in Appendix D.

File information

For every file, main file and supplements, KB needs to know the file name of a zipfile, file name of a main file, file format and file version. Regarding file format KB advises to handle PUID (Persistent Unique Identifier) as a standard. File information can be added in the meta-data and is important for migrating files in the future, which might otherwise become un-readable. By migrating files the digital preservation will be guaranteed for a longer period. For more detailed information read Appendix B: *Background information about the OAIS Model.*

Appendix C. KB use of OAIS Model

The LTP Depot was built in conformance with the Reference Model for an Open Archival Information System (OAIS)[1]. In OAIS terminology the system receives its input from the Producer, which is PDD. In Figure 1 the grey/green box is the LTP Depot. The producer that delivers the objects to the LTP Depot is on the left hand-side of the box. As can be seen the interface between the producer and the archive is the SIP (Submission Information Package)[2].

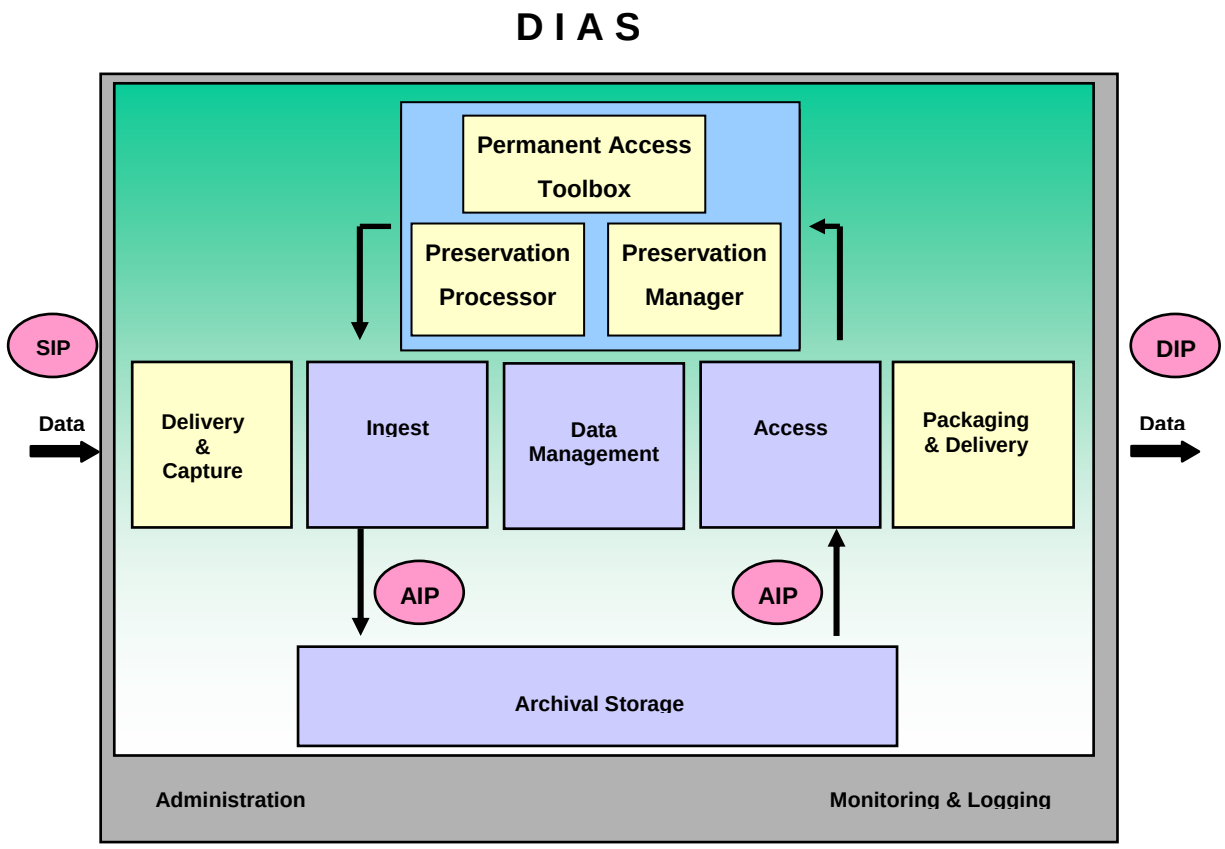

Figure 7: OAIS model (DIAS)

The SIP is the Information Package that is used to ingest the object in the e-Depot. Other Information Packages (the AIP[3] and DIP[4]) as well. According to the OAIS Reference Model each information package (Figure 2) consists of the Content Information (a.k.a. the Information object) together with its Preservation Description Information. The Content Information consists of a Data Object combined with its Representation Information. In Chapter 3 the PDI is explained in detail regarding the different metadata categories.

---

1 http://public.ccsds.org/publications/archive/650x0b1.pdf
2 SIP = An Information Package that is delivered by the Producer (PDD) to the OAIS for use in the Construction of one or more AIPs.
3 AIP = Archival Information Package = the information package as it preserved in the OAIS
4 DIP= Dissemination Information Package = The Information Package, derived from one or more AIPs, received by the Consumer in response to a request to the OAIS.

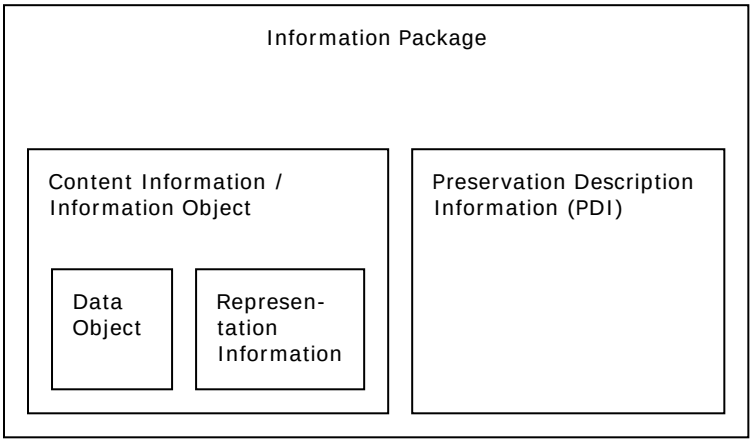

Figure 8: Structure of an Information Package

The Data Object (DO) is the object that together with associated Representation Information is the original target of preservation. It is for example the actual byte stream of a PDF file. To ensure the long-term accessibility of this object the KB does not only require all kinds of metadata and information on the representation of the object, the KB would also prefer to receive the object itself in a durable format. To do so the KB advises on the use of certain file formats and gives guidelines on how to deal with optional settings and choices that can be made within the file format.

Preferred format for structured text documents:

PDF/A

PDF adhering to the KB's PDF Guidelines

The Representation Information (RI) is the information that maps a data object into more meaningful concepts. The RI gives all information that is needed to interpret the DO itself. RI could be the file format specifications of the format in which the data object is coded. The RI can hold the specifications itself or refer to a trusted external registry that holds this information using a persistent identifier to link to the exact specifications within the registry.

Other information that could be seen as RI is the information that describes what software was used to create the DO and what software was the intended software to read the DO. Furthermore all technical dependencies that are needed to run the software (such as hardware, middleware, fonts, external API's etc.) must be described. The KB has a system called the Preservation Manager that holds this information for the KB.

Appendix D. NLM DTD analysis

The National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM) created the Journal Archiving and Interchange Tag Suite with the intent of providing a common format in which publishers and archives can exchange journal content. The Suite provides a set of XML schema modules that define elements and attributes for describing the textual and graphical content of journal articles as well as some non-article material.

The Suite has been written as a set of XML schema modules, each of which is a separate physical file. No module is an entire schema by itself, but these modules can be combined into a number of different schemas.

The Suite can be used to construct schemas for authoring and archiving journal articles as well as transferring journal articles from publishers to archives and between archives.

NCBI/NLM has created several distinct Tag Sets from the Suite of Modules, each with its own purpose. A brief overview of each Tag Set is provided below.

Archiving and Interchange Tag Set	Created to enable an archive to capture as many of the structural and semantic components of existing printed and tagged journal material as conveniently as possible, with no effort made to model any particular sequence or textual format
Journal Publishing Tag Set	Optimized for the archives that wish to regularize and control their content, not to accept the sequence and arrangement presented to them by any particular publisher
Article Authoring Tag Set	Designed for authoring new journal articles, where regularization and control of content is important
NCBI Book Tag Set	Written specifically to describe volumes for the NCBI online libraries

Table 4: Tag Set Overview

Journal Archiving Tag Set (Version 3)

Explanation of this Tag Set[1]

The intent of the Archiving Tag Set is to “preserve the intellectual content of journals independent of the form in which that content was originally delivered”. This Tag Set enables an archive to capture structural and semantic components of existing material without modelling any particular sequence or textual format.

1 http://dtd.nlm.nih.gov/archiving/tag-library/

It was planned that Archiving could be used for conversion from a variety of journal source Tag Sets, with the intent of providing a single format:

- in which publishers could deliver their content to a wide range of archives, and
- into which archives could conveniently translate content from many publishers.

The Archiving Tag Set has a distinct focus on conversion from multiple sources. That focus has made this Tag Set a large and inclusive one. Many elements have been created explicitly so that information tagged by publishers would not be discarded when they converted material from another Tag Set to this one (or one created from this Suite). Care has also been taken to provide several mechanisms (frequently, information classing attributes) to preserve the intellectual content of a document structure when that structure is converted from another Tag Set or schema to this one, even when there is no exact element equivalent of the structure. By design, this is a model for journal articles, such as the typical research article found in an STM journal, and not a model for complete journals. This Tag Set does not include an overarching model for a collection of articles. The exact replication of the look and feel of any particular journal has not been a consideration. Therefore, many purely formatting mechanisms have not been included. At the same time, Archiving is intended to preserve observed content, without resorting to stylesheets or generation of textual elements. For that reason, labels, numbers, and symbols of tables, figures, sidebars, and the like can be recorded as elements, as can the punctuation and spaces inside bibliographic references and lists.

The Journal Archiving Tag Set defines a document that is a top-level component of a journal such as an article, a book or product review, or a letter to the editor. Each such document is composed of one or more parts; if there is more than one part, they must appear in the following order:

- **Front matter (required).** The article front matter contains the metadata for the article (also called article header information), for example, the article title, the journal in which it appears, the date and issue of publication for that issue of that journal, a copyright statement, etc. This is not textual front matter as appears in books, rather this is bibliographic information about the article and the journal in which it was published.
- *Body of the article (optional).* The body of the article is the main textual and graphic content of the article. This usually consists of paragraphs and sections, which may themselves contain figures, tables, sidebars (boxed text), etc. The body of the article is optional to accommodate those repositories that just keep article header information and do not tag the textual content.
- *Back matter for the article (optional).* If present, the article back matter contains information that is ancillary to the main text, such as a glossary, appendix, or list of cited references.
- *Floating Material (optional).* A publisher may choose to place all the floating objects in an article and its back matter (such as tables, figures, boxed text sidebars, etc.) into a separate container element outside the narrative flow for convenience of processing.
- Following the front, body, back, and floating material, there may be either one or more responses to the article or one or more subordinate articles:
    - *Response*. A response is a commentary on the article itself, for example, an opinion from an editor on the importance of the article or a reply from the original author to a letter concerning his article.
    - *Sub-article.* A sub-article is a small article that is completely contained inside another article.

Required elements & attributes

- Top level element: Article (required)
    - **Front Matter structures: Front (required)**
        - Article-meta
        - Contrib(utor)-group
            - Contrib
            - Address
            - Aff
            - Author-comment
                - ✓ P
            - Bio
            - Email
            - Etal
            - Ext-link
            - F(oot)n(ote)
                - ✓ P
            - On-behalf-of
            - Role
            - Uri
            - Xref
            - X
        - Notes
    - *Body and Section structures (optional)*
    - *Back Matter structures (optional)*
        - Ack(nowledgements)
        - App(endix)-group
        - App(endix)
        - Fn-group
            - Fn
                - ✓ P
            - X
        - Glossary
        - Ref(erence)-list
        - Ref(erence Item)
            - Element-citation
            - Mixed-citation
            - Nlm-citation
            - Note
                - ✓ P
                - ✓ Product
            - X
    - *Floating Object structures (optional)*
    - *Sub-article and Response structures (optional)*
        - Front-stub
    - *Block structures (optional)*
        - Array
            - Tbody
                - ✓ Tr

        - Th
        - Td
    - Boxed-text
    - Chem(ical)-struct(ure)-wrap(per)
        - Alternatives
        - Chem-struct
        - Graphic
        - Media
        - Preformat
        - Textual-form
    - Fig(ure)
    - Fig-group
    - Supplementary-material
    - Table-wrap
    - Table-wrap-foot(er)
        - P
        - Fn-group
        - Fn
            - ✓ P
        - Attrib(ute)
        - Permissions
        - X
    - Table-wrap-group
        - Table-wrap
    - Disp(lay)-formula-group
    - Def(inition)-list
    - List
        - List-item
            - ✓ P
            - ✓ Def-list
            - ✓ List
        - X
    - Disp(layed)-quote
    - Speech
        - Speaker
        - P
    - Statement
        - P
    - Verse-group
        - Verse-line
        - Verse-group

Journal Publishing Tag Set (v3)

Explanation of this Tag Set[1]

Publishing is a moderately prescriptive Tag Set, optimized for archives who wish to regularize and control their content, not to accept the sequence and arrangement presented to them by any particular publisher. The Tag Set is also intended for use by publishers for the initial XML tagging of journal material, usually as converted from an authoring form like Microsoft Word. Because Publishing is optimized for regularizing an archive or establishing a sequence of elements to aid print and web production, the Tag Set is smaller than the

1 http://dtd.nlm.nih.gov/publishing/tag-library/

Archiving Tag Set. There are fewer elements, fewer choices in many contexts, and particular element sequence is imposed more often. By design, this is a model for journal articles, such as the typical research article found in an STM journal, and not a model for complete journals. This Tag Set does not include an overarching model for a collection of articles.

The Journal Publishing Tag Set defines a document that is a top-level component of a journal such as an article, a book or product review, or a letter to the editor. Each such document is composed of one or more parts; if there is more than one part, they must appear in the following order:

- **Front matter (required)**. The article front matter contains the metadata for the article (also called article header information), for example, the article title, the journal in which it appears, the date and issue of publication for that issue of that journal, a copyright statement, etc. This is not textual front matter as appears in books, rather this is bibliographic information about the article and the journal in which it was published.
- *Body of the article (optional)*. The body of the article is the main textual and graphic content of the article. This usually consists of paragraphs and sections, which may themselves contain figures, tables, sidebars (boxed text), etc. The body of the article is optional to accommodate those repositories that just keep article header information and do not tag the textual content.
- *Back matter for the article (optional).* If present, the article back matter contains information that is ancillary to the main text, such as a glossary, appendix, or list of cited references.
- *Floating Material (optional).* A publisher may choose to place all the floating objects in an article and its back matter (such as tables, figures, boxed text sidebars, etc.) into a separate container element outside the narrative flow for convenience of processing.
- Following the front, body, back, and floating material, there may be either one or more responses to the article or one or more subordinate articles:
    - *Response*. A response is a commentary on the article itself, for example, an opinion from an editor on the importance of the article or a reply from the original author to a letter concerning his article.
    - *Sub-article.* A sub-article is a small article that is completely contained inside another article.

Required elements & attributes

- Top level element: Article (required)
    - **Front Matter structures: Front (required)**
        - Journal-meta
            - ✓ Journal-id
            - ✓ Issn
        - Article-meta
            - ✓ Title-group
                - ➢ Article-title
            - ✓ Pub-date
                - ➢ Year
            - ✓ Fpage
        - Contrib-group
            - ✓ Contrib
                - ➢ Name
                    - ❖ Surname
        - Bio

- *Body and Section structures (optional)*
- *Back Matter structures (optional)*
  - Ack(nowledgements)
  - App(endix)-group
  - App(endix)
    - Label
    - Title
  - Glossary
  - Ref(erence)-list
  - Nlm-citation
- *Floating Object structures (optional)*
- *Sub-article and Response structures (optional)*
  - Front stub
- *Block structures (optional)*

See Journal Archiving Tag Set.

Article Authoring Tag Set (v3)

Explanation of this Tag Set[1]

The Article Authoring Tag Set creates a standardized format for new journal articles that can be used by authors to submit publications to journals and to archives such as PubMed Central. While in theory the document scope is the same as for the Publishing Tag Set, in practice Authoring defines elements and attributes that describe the content of typical research-style journal articles.

This is a Tag Set optimized for authorship of new journal articles, where regularization and control of content is important, and where it is useful rather than harmful to have only one way to tag a structure. Therefore, Authoring is more *prescriptive* than *descriptive* and includes many elements whose content must occur in a specified order.

Since an author is assumed to be creating and submitting an article for submission to a journal or journals, no publishing history or journal-specific information has been included in this Authoring Tag Set.

By design, this is a model for journal articles, such as the typical research article found in an STM journal, and not a model for complete journals. This Tag Set does not include an overarching model for a collection of articles.

The Article Authoring Tag Set defines a document that is a top-level component of a journal such as an article, a book or product review, or a letter to the editor. Each such document is composed of one or more parts; if there is more than one part, they must appear in the following order:

- **Front matter (required).** The article front matter contains the metadata for the article (also called article header information), for example, the article title, the names of the contributor(s), and the abstract. This is not textual front matter as appears in books, rather this is bibliographic information about the article.
- **Body of the article (required).** The body of the article is the main textual and graphic content of the article. This usually consists of paragraphs and sections, which may themselves contain figures, tables, sidebars (boxed text), etc.

1 http://dtd.nlm.nih.gov/articleauthoring/tag-library/

- *Back matter for the article (optional).* If present, the article back matter contains information that is ancillary to the main text, such as a glossary, appendix, or list of cited references.

Required elements & attributes

- Top level element: Article (required)
    - **Front Matter structures: Front (required)**
        - Article-meta
            - ✓ Title-group
                - ➢ Article-title
            - ✓ Contrib-group
                - ➢ Contrib
                    - ❖ Anonymous
                    - ❖ Collab
                    - ❖ Name
                        - Surname
        - ✓ Abstract
    - **Body and Section structures (required)**
        - Body
    - *Back Matter structures (optional)*
        - Nlm-citation
    - *Block structures (optional)*

    See Journal Archiving Tag Set.

NCBI Book Tag Set (v3)

This Tag Set is not described in this document, because the PEER project only concentrates on articles (Stage-2 content).

Explanation of this Tag Set[1]

The NCBI Book Tag Sets were written using the Publishing Tag Set as a base and adding book-specific elements. The Book Tag Sets define elements and attributes that describe the content of books (such as pamphlets and monographs) and book collections, respectively.

Unlike the Journal Tag Sets, which were written generically to be of wide-ranging applicability beyond NLM, the Book Tag Sets were written with a more modest purpose, to describe volumes for the NCBI online libraries.

The book modules have been arranged to form two Tag Sets:

- the NCBI Book Tag Set; and
- the NCBI Book Collection Tag Set.

The NCBI Book Tag Set, while written to describe both the metadata for a book and the content of the book, can also be used to describe only the metadata for both books and book parts, such as “Chapters”. The NCBI Collection Tag Set describes a document that contains metadata for a grouping or “collection” of books, potentially followed by textual information about the books in the collection, and a listing of the books; the content of the books is not included in the Collection Tag Set.

1 http://dtd.nlm.nih.gov/book/tag-library/

Appendix E. PEER DTD

PEER Metadata	NLM Example[1]	Explanation[2]
Journal title	<journal-title>	
Article ID/DOI	<article-id>	The @pub-id-type attribute may be used to name the type of identifier (such as DOI or SICI), or the organization or system (such as PubMed) that defined this identifier. This attribute need only be used if the type is known, for example, to identify DOIs explicitly. DOI: To indicate relations betweens objects, for example a stage-2 manuscript with its stage-3 published article.
Article title	<article-title>	
Journal abbreviation/code	<abbrev-journal-title>	
Volume number	<volume>	
Issue number	<issue>	
Author	<name><surname><given-names><suffix>	
Author email	<email>	
Author country/coverage	<corresp> or <country>	Information concerning which of the authors (or other contributors) is the corresponding contributor, to whom information requests should be addressed.
Affiliation	<aff>	
Publisher name	<publisher-name>	

1 Tags named according to the Journal Publishing Tag Set NLM DTD.
2 http://dtd.nlm.nih.gov/publishing/tag-library/

Publisher ID		Is probably not present in Journal Publishing Tag Set.
Publishers' location	<publisher-loc>	
eISSN	<issn>	
Year of acceptance	<date date type=”accepted”><year>	History container
Month of acceptance	<date date type=”accepted”> <month>	History container
Day of acceptance	<date date type=”accepted”><day>	History container
Year of publication	<pub-date pub-type=”pub”><year>	
Month of publication	<pub-date pub-type=”pub”> <month>	
Day of publication	<pub-date pub-type=”pub”><day>	
Keyword	<kwd>	
Abstract of the article	<abstract>	
Embargo		Is probably not present in Journal Publishing Tag Set.
Page range	<page-range>	First and last page.
Language article		Default = English Is probably not present in Journal Publishing Tag Set.
Copyright	<copyright-holder>	Permissions container
Type-description (category/theme of the publication)	<article-type>	Default = article
Availability (restricted/open access)	<open-access>	Restricted/open access
Productiondate-key	<date in citation type=update>	‘Update tag’ Publishers will deliver updates of metadata of stage-2 manuscripts.
File names	<related-object>	Container element for a text link to a published

		related object other than a journal article, possibly accompanied by a very brief description of the object. For example, the related object might be a related book, a chapter in a book, or a figure or graphic from another published source (supplements) which should be repeatable.
File formats		
File versions		

Table 5: PEER DTD

Appendix F. KB DTD

To ensure long-term preservation and render permanently access to scientific content KB uses the following metadata elements:

- Journal title (mandatory)
- Journal abbreviation/code
- Volume number (mandatory; hierarchy function catalogue system)
- Issue number (mandatory; hierarchy function catalogue system)
- Article ID (mandatory)
- DOI (preferred; mandatory)
- Article title (mandatory)
- Publisher name (mandatory)
- Publishers' location
- Copyright
- Author ID (mandatory)
- e-ISSN (preferred; mandatory)
- Print-ISSN
- Pages range (mandatory)
- Year of publication (mandatory)[1]
- Month of publication
- Day of publication
- Abstract of the article
- Availability
- Language article
- Keyword (mandatory)
- Type-description (category/theme of the publication)
- Productiondate-key
- Embargo
- File name
- File format
- File version
- Supplements
- Relation

1 Usage of dates is variable. Accepted date, by year, month, day, may be used instead of the publication date, because stage-2 manuscripts are published in this stage.

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

PubMan@MPDL 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 that focuses on publication management. PubMan is used at present by several early-adopter Max-Planck Institutes. It is anticipated to be extended to all institutes within the Max-Planck Society (MPG) and to replace the current eDoc publication repository of MPG.

PubMan@MPDL offers at present the following general usage statistics and reports:

- Numbers of retrievals for a specific item by users (anonymous/all)
- Numbers of file downloads for a specific item by users (anonymous/all)
- Numbers of downloads for a specific file by users (anonymous/all)

These statistics are aggregated on item/file level and are as such available in XML format via web service. Raw statistics can be delivered but not via a web service.

Examples of statistical reports:

Example 1: Number of retrievals for a specific item by all/anonymous users:

```
<report xmlns="http://www.escidoc.de/schemas/report/0.3">
   <report-definition objid="86"/>
   <report-record>
      <parameter name="itemid">
         <stringvalue>escidoc:1641:5</stringvalue>
      </parameter>
      <parameter name="itemrequests">
         <decimalvalue>2</decimalvalue>
      </parameter>
   </report-record>
</report>
```

Example 2: Number of all file downloads of an item by all/anonymous users:

```
<report xmlns="http://www.escidoc.de/schemas/report/0.3">
   <report-definition objid="90"/>
   <report-record>
      <parameter name="itemid">
         <stringvalue>escidoc:1641</stringvalue>
      </parameter>
      <parameter name="filerequests">
         <decimalvalue>2</decimalvalue>
      </parameter>
   </report-record>
</report>
```

Example 3: Number of file downloads of specific file of an item by all/anonymous users:

```
<report xmlns="http://www.escidoc.de/schemas/report/0.3">
   <report-definition objid="88"/>
   <report-record>
      <parameter name="fileid">
         <stringvalue>escidoc:1642</stringvalue>
      </parameter>
      <parameter name="filerequests">
         <decimalvalue>9</decimalvalue>
      </parameter>
   </report-record>
</report>
```

All reports above are generated for a particular timeframe (report parameter).

Examples of aggregated data:

Aggregated data are basis for reports delivery and they are stored in a relational database. The granularity of aggregated data simply depends on the aggregation definition. For statistics from Example 1 (number of retrievals of an item by registered users) PubMan repository provides the following aggregated data for each item:

Operation: retrieval
ObjectId: escidoc:39207
User Id: escidoc:22127
Year: 2008;
Month: 11;
Number of retrievals: 6

Examples of raw statistical data:

Each aggregated statistic report (described in examples 1−3) is derived from raw data stored in a relational database and that have the following information:

Timestamp − (i.e. the timestamp, including timezone when retrieval of an object occurred) e.g. "2008-11-12 15:05:21.639"

XML data − Information on the retrieval operation and object Id (in XML format) e.g.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<statistic-record>
   <scope objid="1"/>
   <parameter name="handler">
      <stringvalue>de.escidoc.core.om.service.ItemHandler</stringvalue>
   </parameter>
   <parameter name="request">
      <stringvalue>retrieve</stringvalue>
   </parameter>
   <parameter name="interface">
      <stringvalue>SOAP</stringvalue>
   </parameter>
   <parameter name="internal">
      <stringvalue>0</stringvalue>
   </parameter>
   <parameter name="object_id">
      <stringvalue>escidoc:8123</stringvalue>
   </parameter>
   <parameter name="successful">
```

```
      <stringvalue>1</stringvalue>
   </parameter>
   <parameter name="elapsed_time">
      <stringvalue>173</stringvalue>
   </parameter>
   <parameter name="user_id">
      <stringvalue>escidoc:user42</stringvalue>
   </parameter>
</statistic-record>
```

In the example above, the raw statistical record contains information that an item with id: escidoc.8123 has been retrieved successfully by a certain user account and via the SOAP interface of the Item service.

At the time of writing, further information on current practice in the provision of usage data in other participating repositories was not readily available. By February 2009 MPDL will gather details as set out below, and extend the PubMan statistics with more precise data on events for items and files associated to an item, such as: IP address, Open Access/Not Open Access, event (retrieval, export, update), etc.

INRIA
HAL uses Apache and produces logfiles in the format NCSA combined.

(UniBi) http://repositories.ub.uni-bielefeld.de/biprints/
BiPrints uses APACHE and produces logfiles 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)"
```