

PubMan – one Repository with multiple Usage und Re-Use Possibilities

Juliane Müller

Max Planck Digital Library, Research & Development,
Amalienstraße 33, 80799 Munich, Germany
jmueler@mpdl.mpg.de

ABSTRACT

PubMan is an application which allows members of research organizations to store, manage and enrich their publications. The app is based on the eSciDoc infrastructure, a joint project run by the Max Planck Digital Library (MPDL) and the Fachinformationszentrum (FIZ) Karlsruhe.

Presenting scholarly work in the World Wide Web has become an important and common procedure for research communities seeking to enhance the visibility of their research results as well as to initiate scientific collaboration and information exchange. In response to that trend much emphasis has been put on the possibility of providing multiple re-use options for metadata, full texts and supplementary material during the conception and development of PubMan. Our repository software facilitates the integration of user-defined publication lists in local websites as well as in personal and topic-centered WordPress blogs. The paper will depict these two re-use possibilities with examples of operational usage scenarios after giving an overview of the basic concepts and functionalities of PubMan.

KEYWORDS

Digital Archive, eSciDoc, Institutional Repository, Max Planck Society, Publication Management, Publication Repository, Re-Use Possibilities, Web 2.0

1. INTRODUCTION

In the context of the eSciDoc project (<http://www.escidoc.org>), the Department of Research& Development at the MPDL is responsible for the development of PubMan (<http://pubman.mpdl.mpg.de/>) and user support at the Max Planck Institutes. The aim of the eSciDoc project was to develop a stable eScience infrastructure [1] for sustainable access to the research results and research data of the Max Planck Society (MPS) as well as to build an interdisciplinary scholarly communication platform for scientific collaboration. Thus it was a considerable challenge for the project partners to establish an infrastructure able to deal with the variety of disciplines and the multiplicity of data formats and information entities required to address different research scenarios. Furthermore the proposed research environment should allow for the dissemination, re-use and mash-up of the stored data in multiple ways [2] [3]. As eSciDoc is an open source development, it should be possible for any interested research organization to re-use it.

Based on the eSciDoc infrastructure, PubMan is one of currently three applications developed by the MPDL. FACES [4] is a web-based app for the collection of image data with corresponding metadata for each picture. ViRR (Virtueller Raum Reichsrecht) [5] deals with digitized artefacts and provides a structured navigation through the scanned text resources. PubMan [6] [7] is the future institutional repository of the MPS and digital archive for publications and supplementary material. The development of PubMan is an ongoing process and addresses the heterogeneous requirements of scholars, librarians and IT staff of the MPS. The institute's needs are characterized by configurable publication workflows, multiple features for data deposit and data management as well as appropriate ways of presenting the stored material. Various usage scenarios had already been identified and specified through a close cooperation between early-adopter institutes and the MPDL staff. A first fully-operational release of PubMan was deployed in May 2009. As PubMan is to be the future repository of the MPS the institutes are currently supported by the MPDL while working on the migration of their data from the former repository eDoc (<http://edoc.mpg.de/>) to the new system [8]. PubMan is an open source application - about 15 national and international institutions are currently evaluating the re-use of PubMan both in its own right and as a basis for further collective development.

2. THE ESCIDOC INFRASTRUCTURE

The eSciDoc infrastructure is designed as a service-oriented architecture (SOA) [9] [10], based on a Fedora Commons platform (<http://www.fedora-commons.org>). Application and discipline-specific solutions can be built on top of the infrastructure. It is not a monolithic system and could be better described as a set of loosely-connected services which can be implemented easily and independently. The eSciDoc services are grouped into three service layers: core services, intermediate services, and application services [11]. As PubMan is fully embedded in the eSciDoc infrastructure it is able to profit from all existing eSciDoc services, e.g. persistent identification, automatic versioning, digital long-term preservation, validating service, transformation, data acquisition as well as the support of standardized metadata profiles (Qualified Dublin Core). Furthermore eSciDoc provides standardized interfaces such as OAI-PMH, SWORD, SOAP, RSS or REST.

3. PUBMAN AS A PUBLICATION REPOSITORY

Different publication types (e.g. journal article, book chapter, proceedings paper, poster, talk, thesis ...) can be deposited in short or detailed genre-specific submission masks. The PubMan user is able to provide copyright information and Creative Common licenses (<http://creativecommons.org>) to uploaded full texts and supplementary material. Furthermore three different visibility levels - public, private and restricted to a specific user group - can be assigned to full texts and related material. For the last two access levels it is possible to provide an embargo date. Publications can also be submitted by fetching metadata and full texts from other sources (arXiv, BioMed Central, Spires, PubMed Central) as well as by running a multiple import from EndNote, RIS, WoS or BibTex.

PubMan supports a quick and an advanced search. Full-text searching is offered; respective search terms are highlighted in the results list. With individual filter and sorting mechanisms the user is able to prepare a specific list of publication items for export either as EndNote, BibTex or XML files. Moreover single items or item lists can be exported as RTF, ODT, PDF, HTML or snippet format (XML) in a particular citation style (APA and AJP are currently implemented). All export options are available for download or via email.

Statistical analyses with respective visualizations will be generated on accessed metadata and full texts both for logged-in and anonymous users [12].

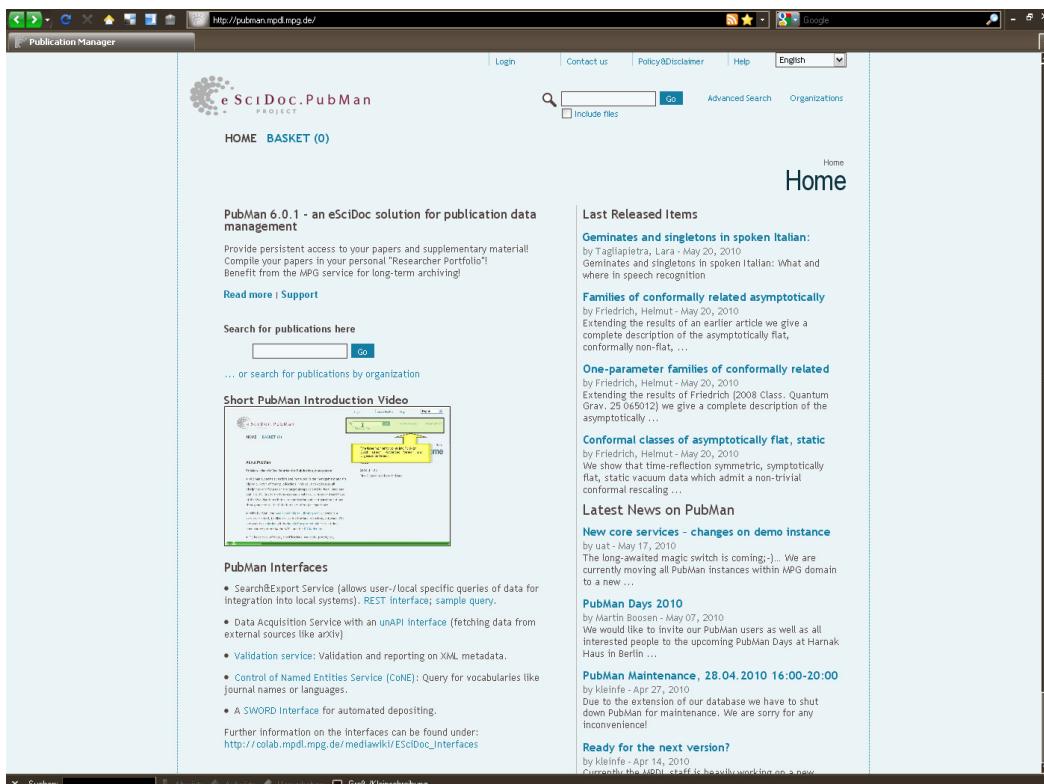


Figure 1: Screenshot of the PubMan Homepage

All metadata records as well as each attached file are persistently identified. So far support of the Handle System (<http://www.handle.net>) has been implemented. As soon as any metadata or the numbers of attached full texts are changed a new version will be generated automatically. When accessing a particular publication item in PubMan the most recent version is displayed. All released versions of the publication item are visible and retrievable for every user; additionally all unreleased versions of the publication item are visible and retrievable for the owner and moderator of the respective publication item.

So far two publication workflows have been implemented: a simple workflow offering easy release of the publication by the depositor as well as a more complex workflow which comprises various stages of quality assurance management before a publication may be released in the repository.

The integration of the eSciDoc CoNE service (Service for Control of named Entities) [13], a separate web service for the administration of controlled vocabulary, provides the user with autosuggest list support during submission when entering persons, journals, Creative Common licenses, mime types, languages and DDC subjects (<http://www.oclc.org/dewey>). Soon the service will be further extended by other subject classifications like PACS (<http://www.aip.org/pacs/>) or institute-specific categorizations. The CoNE service facilitates the assignment of a unique ID to every individual person (e.g. author, editor etc.), allowing the tracking of all name variants and historic relations of a person to certain Organizational Units (OUs). OUs are administered in a similar way – unique IDs are assigned to them as soon as they are created within the PubMan environment. This means it is possible to search for specific OUs, departments or projects. Due to the information stored within the CoNE service for persons PubMan is able to automatically generate an individual researcher portfolio for every author. Besides information about the OU the author is affiliated to or research fields he or she is interested in, this researcher portfolio comprises a complete publication list of all publications archived in PubMan [14].

The focus of the future repository of the MPS is on the publication with its bibliographic metadata and the possibility of providing open or alternatively controlled access to the stored full texts. Furthermore PubMan enables the addition of supplementary material related to the submitted publication. This kind of research data (not envisaged to be in the terabyte dimension) could be available in flexible electronic formats (for example Excel, SPSS, image, audio or video files etc.). The possibility of storing such data together with the actual publication enables scholars and research communities to reproduce scientific results and to re-use the provided supplementary material for further research questions.

4. PUBMAN – ONE SOURCE WITH MULTIPLE RE-USE OPTIONS

Presenting research output on the World Wide Web has become an important and common procedure amongst scholar communities aiming to enhance scientific collaboration and information exchange. Thus scholars, research projects and entire research institutions are interested in making all their relevant research results accessible on their respective websites. In response to the multifarious requirements of the MPS stakeholders within the fields of research, library and IT much emphasis has been put on the possibility of providing various re-use options for metadata, full texts and supplementary material during the conception and development of PubMan.

The implementation of a standardized interface (REST) for Search and Export has made it possible to generate queries of data in specific citation styles, sorting options and formats [15]. This functionality supports the compilation of complex reports as well as the integration of this kind of preformatted publication list in local websites. Moreover the implementation of a particular plug-in for the blog software WordPress (<http://wordpress.org/>) enables the dynamic embedding of PubMan publications within both personal and topic-centered blogs. The following two paragraphs will describe operational usage scenarios for the re-use of metadata and components which are stored in PubMan.

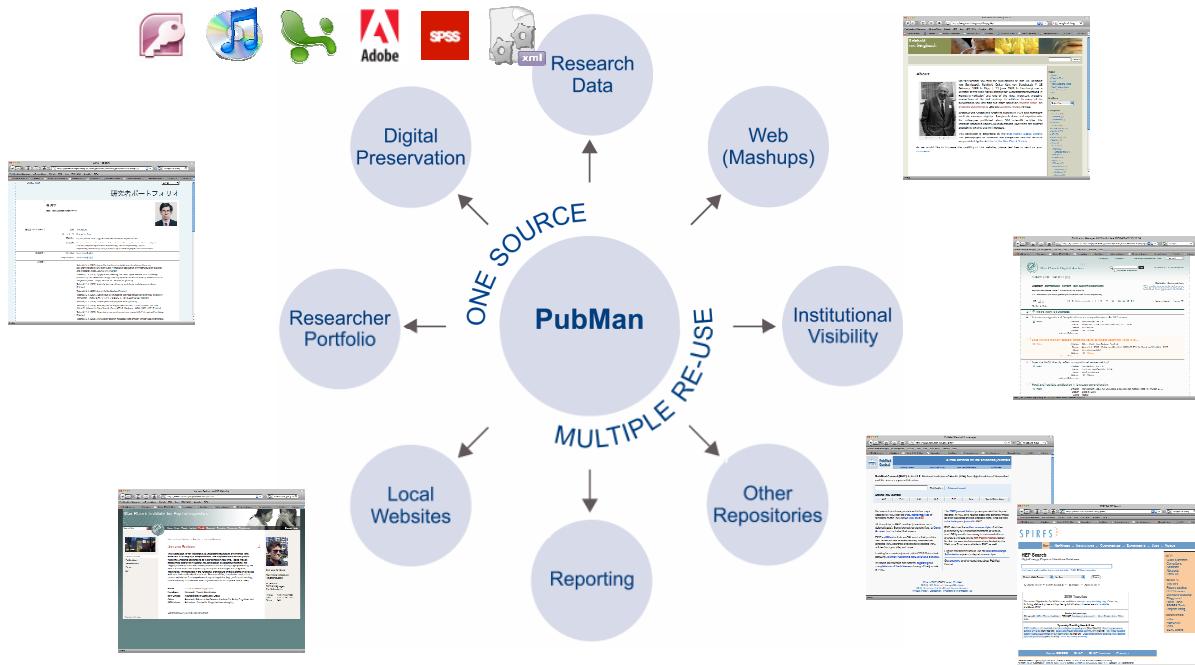


Figure 2: Visualization of various PubMan re-use options

4.1 THE INTEGRATION OF PUBMAN DATA INTO LOCAL WEBSITES

The MPI for Psycholinguistics (MPI PL) in Nijmegen, Netherlands (<http://www.mpi.nl>) has worked with PubMan since April 2009. Their desire to integrate publication lists with open access full texts and supplementary material on different levels on their institute's website challenged them to tackle a complete redesign of their webpage [16].

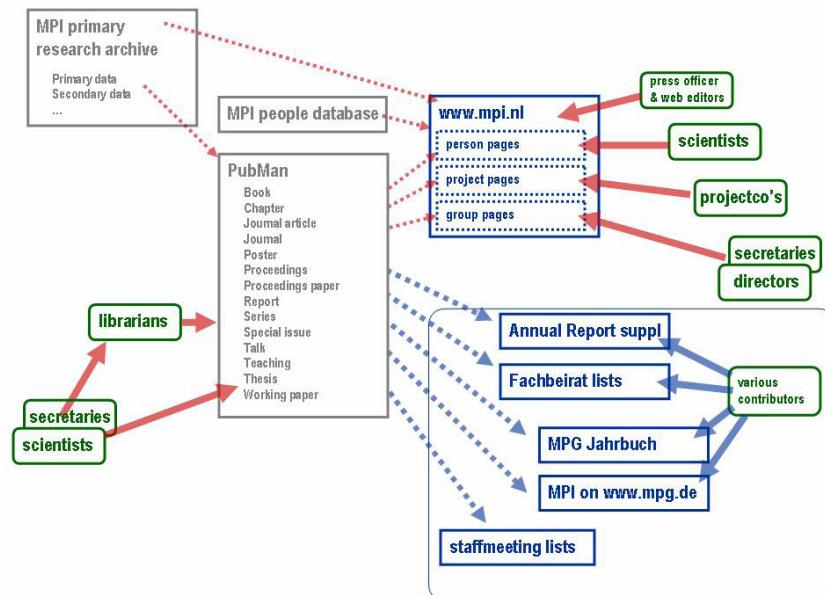


Figure 3: Visualization of the MPI PL information flow¹

¹ I would like to thank Jos van Berkum (MPI PL, Nijmegen, Netherlands) for providing that figure.

The MPI PL decided to combine their content management system (CMS) Plone (<http://plone.org>), which forms the technical infrastructure of their website, with the institutional repository PubMan, where all their publications are stored. Figure 3 shows the information flow at the institute. Pages on the institute's website with respective publication lists for persons, projects and groups could be enriched with the MPI PL data stored in PubMan. Scientists, the secretaries of the research groups and the library staff deposit the publications in PubMan. The latter are responsible for the quality assurance of the data. The PubMan submission mask provides a so-called "locator" field which enables linking to external material. The MPI PL maintains a large research data archive (<http://corpus1.mpi.nl>), which comprises open access research data and is browsable and searchable over the web. At the MPI PL the locator option in PubMan is used for linking to respective files in the research data archive of the institute. Of course it is also possible to store supplementary material physically in PubMan. Figure 3 also shows that the PubMan data can not only be used to fill the publication lists on the institute's website but also to generate annual reports or other specific compilations.

The Search and Export REST interface makes it possible for the CMS of the MPI PL to harvest the respective data from the PubMan database. A particular search query, specifying the necessary selection criteria and desired export formats, is sent to the PubMan server overnight and an XML file containing the metadata is returned to the CMS. The MPI PL harvests the data in a "snippet" format, which is an XML file also including the bibliographic citation in the required export style (APA). Full texts and supplementary material are not imported into the CMS; the harvested metadata includes the persistent URLs to the respective resources.

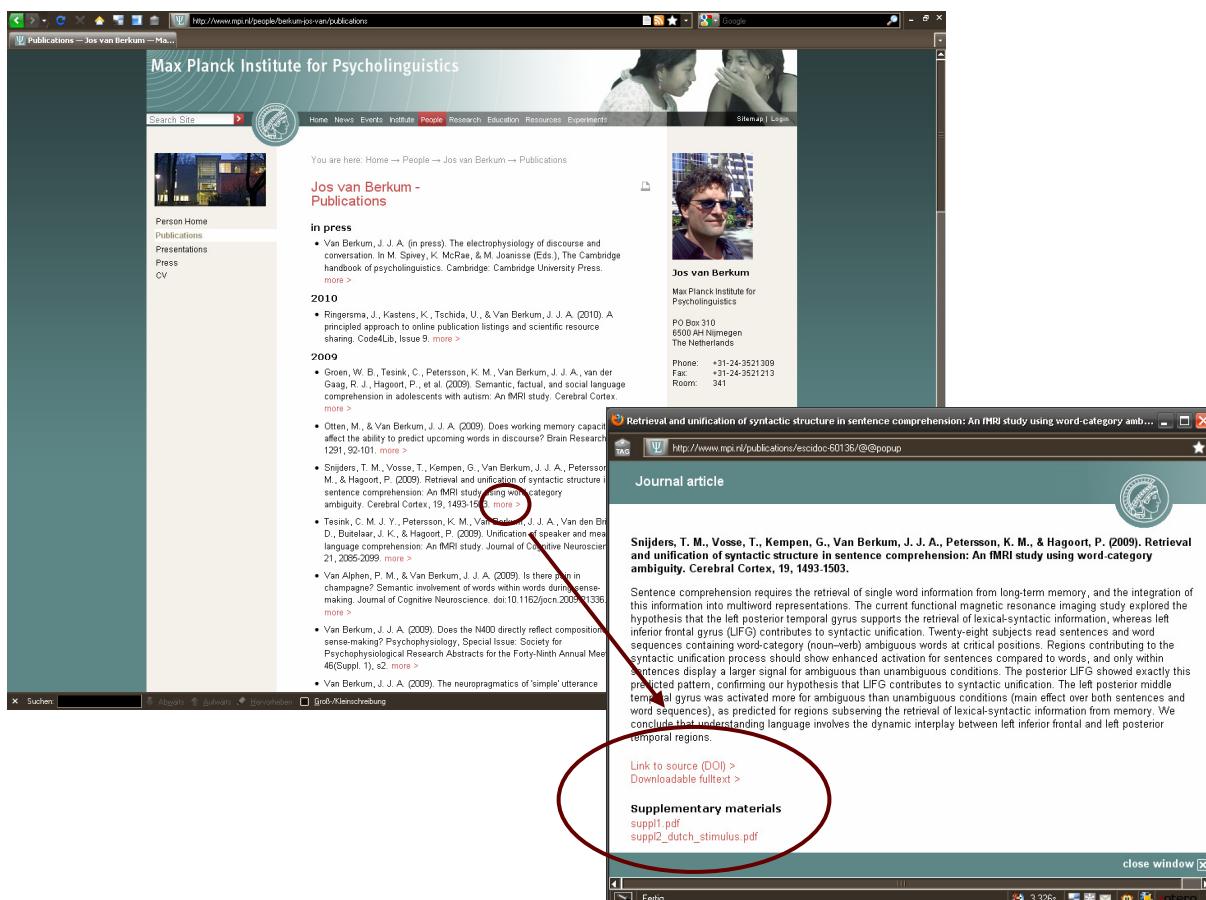


Figure 4: Screenshot of a person page on the MPI PL website with pop-up window for further information about the publication and links to the full text and supplementary material

Figure 4 shows a screenshot of a publication list on a personal page within the institute's website. The references are sorted by particular criteria and are formatted in APA citation style. The sort order of the publications is either publication year or publication type by default, but could be also individually adapted per author and reuse platform. Clicking on the link "more" triggers the opening of a pop-up window. There the APA citation is complemented by an abstract of the publication together with persistent URLs to full texts and supplementary material if open access is provided to them.

Further institutes of the MPS are interested in and are even already working on integrating PubMan data into their institute-specific CMS. The Python script for the re-use in any Plone based website is archived in PubMan [17]. The legal institutes of the MPS in cooperation with the MPDL are currently working on the possibility of feeding PubMan data into the legal institute's webpages which are administered by the CMS software CONTENS (<http://www.contens.de>). Thus the re-use scenarios for feeding local websites are constantly growing.

4.2 THE INTEGRATION OF PUBMAN DATA INTO WORDPRESS BLOGS

The Search and Export REST interface enables PubMan data to be fetched and integrated into WordPress blogs. A plug-in [18] allows blog owners to automatically import new PubMan publication items into their WordPress blog as well as to extract automatic updates of existing publication items. The new publications can be imported as draft or published versions. Similar to the procedure of feeding local websites, a specific Search and Export query has to be generated, comprising the specifications of a particular sorting option together with the required citation style and export format. The WordPress plug-in harvests the respective PubMan data via the REST interface and a preformatted publication list in XML format is returned to the plug-in. Publications can be sorted by different criteria (e.g. free keywords, publication year, genre types, local tags etc.). Every publication is one post displayed in a particular citation style. To get more information about the complete metadata of the publication a link to the PubMan item is added. The full text, which is stored in PubMan, is directly accessible from the blog. The plug-in also creates a RDF presentation of each publication; the respective RDF is linked below every blog post.

Once activated the plug-in automatically – at regular intervals – takes care of harvesting new or updated PubMan publications. In blogs each publication gets its own announcement page. The author is able to personalize these pages, for example by adding an abstract, photos, links or background information. The blog owner is free to adapt the CSS of the blog according to her or his own ideas. The integration of PubMan data into blogs makes it possible to create an individual web presence for presenting research output in a personalized manner. Blogs are of equal interest to individual scientists wanting to provide information about their research fields and scientific results and for entire research groups or communities focusing on their collection of research outputs.



Figure 5: Screenshot of the Reinhold v. Sengbusch Collection – a person centered researcher blog

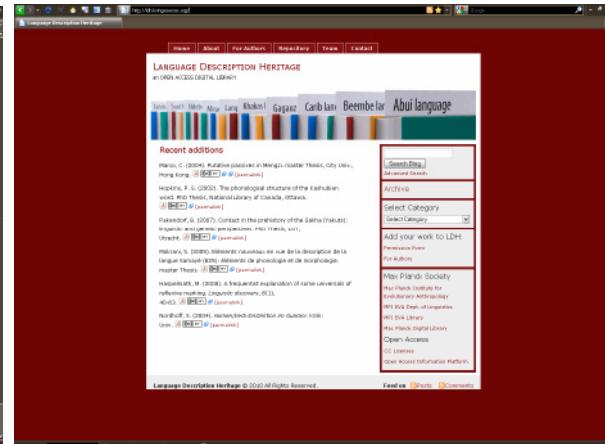


Figure 6: Screenshot of the collection of Language Description Heritage – a research project centered blog

Figure 5 shows the screenshot of the Reinhold von Sengbusch collection (<http://sengbusch.blogs.mpdl.mpg.de>) – an example of a person centered researcher blog. This collection will eventually contain the complete works of Professor Reinhold von Sengbusch, the former director of the “MPI für Kulturpflanzenzüchtung” in Hamburg, Germany. All publications are stored in PubMan. As soon as a new Sengbusch publication becomes available on PubMan, it is automatically fetched by the WordPress plug-in and posted into the blog.

Figure 6 shows a screenshot of a project-centered blog for the collection of “Language Description Heritage” (<http://ldh.blogs.mpdl.mpg.de>) whose aim is to provide easy access to descriptive material about the world’s languages. This collection is being compiled at the MPS as an open access digital repository of existing scientific contribution describing world-wide linguistic diversity.

Both blogs are under construction – they will gradually be enriched by further publications with direct access to full texts. These blogs will enhance the visibility of the research output of the respective scientist or project group. They will also support the overall aim of providing open access to scientific information for everybody.

5. CONCLUSION

Our aim was to develop a repository software offering efficient publication management. PubMan itself is the “invisible” source comprising the publications with their bibliographic metadata, full texts and supplementary material. The support of multiple re-use options was the challenge we wanted to meet. And we are certain PubMan already provides convincing options for the dissemination, re-use and mash-up of the stored data: PubMan users are able to compile complex publication lists, reports and collections. The application furthermore offers good connections to content management systems for the purpose of integrating of PubMan data into local websites. The PubMan plug-in developed for WordPress means that both personal and project-oriented blogs can be automatically enriched by preformatted publication lists.

For a research organization PubMan is both an appropriate digital archive - with its standardized, sustainable and efficient underlying eSciDoc infrastructure - and an institutional showcase for all its research results. Scholars are assessed by their research output. Thus publication lists are of major importance for them. Creating a personal web presence where researchers are able to enrich information about their scholarly work with automatically updated publications lists is a great opportunity for them to enhance the visibility of their scientific output. Publications will be found more easily by other researchers – and this is the basis for the initiation of scientific discourse and information exchange which in turn leads to further research questions.

Finally, PubMan as a single source combining data storage and the means of making publications visible in broader contexts and with a wide range of purposes is also a huge incentive for open access.

6. REFERENCES

- [1] Dreyer, M. et al. 2007. eSciDoc – a scholarly information and communication platform for the Max Planck Society. Conference Paper for the GES 2007, Baden-Baden, Germany. <http://www.ges2007.de/papers/>
- [2] Bulatovic, N., Tschida, U., Gros, A. 2008. eSciDoc - a service infrastructure for management of Cultural Heritage content. Conference Paper for the Conference on Virtual Systems and MultiMedia dedicated to Digital Heritage, Limassol, Cyprus. <http://edoc.mpg.de/get.epl?fid=48783&did=367778&ver=0>
- [3] Dreyer, M. & Tschida, U. 2009. eSciDoc – Das Repository-Konzept der Max Planck Digital Library. *cms-journal*. Volume 32, pp. 56-59. Humboldt University of Berlin.
<http://edoc.hu-berlin.de/cmsj/32/dreyer-malte-56/PDF/dreyer.pdf>
- [4] More information about FACES: <http://colab.mpdl.mpg.de/mediawiki/Faces>
- [5] More information about ViRR: http://colab.mpdl.mpg.de/mediawiki/ViRR:_Virtueller_Raum_Rechtsrecht
- [6] PubMan Portal at the MPDL MediaWiki CoLab: <http://colab.mpdl.mpg.de/mediawiki/Portal:PubMan>
- [7] PubMan Category at the MPDL MediaWiki CoLab: <http://colab.mpdl.mpg.de/mediawiki/Category:PubMan>
- [8] Tschida, U. 2009. PubMan – ein Repository für die MPG. Article for the MPS Yearbook 2010.
<http://edoc.mpg.de/442612>
- [9] Definition of SOA: http://colab.mpdl.mpg.de/mediawiki/Service_Oriented_Architecture
- [10] More information about the eSciDoc SOA: http://colab.mpdl.mpg.de/mediawiki/ESciDoc_SOA
- [11] More information about the eSciDoc services: http://colab.mpdl.mpg.de/mediawiki/ESciDoc_ServiceLayers
- [12] Complete list of all PubMan functionalities which are currently implemented:
http://colab.mpdl.mpg.de/mediawiki/PubMan_Functionalities
- [13] More information about the eSciDoc Service for Control of named Entities:
http://colab.mpdl.mpg.de/mediawiki/Service_for_Control_of_Named_Entities

[14] Example of a researcher portfolio automatically generated by PubMan:
<http://pubman.mpdl.mpg.de/cone/persons/resource/persons1008>

[15] More information about the PubMan REST Interface:
http://pubman.mpdl.mpg.de/search/SearchAndExport_info.jsp

[16] Ringersma, J. et al. 2010. A principled approach to online publication listings and scientific resource sharing. *Code4Lib Journal*. Issue 9. (submitted article).

[17] Python script for the re-use of PubMan data in Plone based websites:
<http://pubman.mpdl.mpg.de/pubman/item/escidoc:101899:8>

[18] Installation description for the WordPress plug-in which enables the possibility to integrate PubMan data into blogs: http://colab.mpdl.mpg.de/mediawiki/PubMan_Wordpress_Plugin

* I would like to thank Janet MacKenzie for assisting with the editorial work.