
PubMan content categories needed for the MPI-PL website, and *initial* specs for website rendering of Pubman content

Feb 4 2008

Jos van Berkum

Rationale

In the January 31 internal PubMan meeting, we converged on a set of genres that would serve the MPI-PL needs (see separate note from Karin), as well as on suitable genre labels for the MPI-PL PubMan interface. However, what still needs to be specified is **what we need in terms of Pubman content categories**. The MPDL proposal listed in Karin's slides (preprint, postprint, publisher version, abstract, table of contents, supplementary materials, copyright agreement, correspondence, any full text) is sensibly oriented towards the canonical PubMan entry, a journal publication. However, there will be many more genres, at least in the PubMan system for our institute.

In the below, I analyze what we'd want to store in a PubMan record *from a website's perspective*. Note that there are other equally valid perspectives (annual report, fachbeirat/staffmeeting listings, Jahrbuch, archiving for strictly internal use, e.g. Institute memory, quality control, etc.) that might require additional content. But this is a good place to start, because our Plone website will be a 'heavy user' of PubMan (and will as such also motivate people to put their stuff in PubMan).

Although originally geared to clarify what we need to store in and retrieve from PubMan, the below also serves as an **initial requirements specification for the PubMan-Plone interface (PPI)**.

For every record genre, I give an example of the website entities in which PubMan-retrieved data of this type play a role, and along the way I discuss requirements for PubMan or the PubMan-Plone interface (PPI). Webcontents will be in arial, and pubman-delivered contents are in orange.

The genre names are those we decided on in the January 31 internal meeting, and my description below follows the list we used in that meeting (made by Nicole Kondic, MPDL). All but the last four genres have a clear mapping to genre types currently used in PubMan, the last four are new (see Karin's memo). I adopt the same strategy as for the genres: let us first think about what we need, using optimal labels for us, and then see how this maps onto existing categories/labels in the current PubMan release.

For each genre, I try to think of additional bits of content that we'd want to pull out of the PubMan record at hand. Note that although I list several such content types per genre, listing them is not a plea to store everything we can possibly think of with a record. Most of the content categories listed are optional, i.e., can be instantiated if somebody wishes to do so; if instantiated, however, the idea is that we can pull them out for website use in a principled way, foreseen in the PPI. At some later time, decisions about what is obligatory and what is optional in PubMan, will need to be made.

For each genre, we also need to think of additional metadata relevant for selecting/sorting records, with respect to the website outlet, but also with respect to the other identified outlets (Annual Report, Jahrbuch, etc).

(1) Journal article

(1.1) The typical website rendering is an APA-style reference/citation with some additional links, as part of a list of many more such citations:

- Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 31(3), 443-467. ([abstract](#) [fulltext](#) [supplementary-materials](#) [external-link](#))

Abstract (always supplied, hence obligatory content in PubMan)

Action: open abstract in a separate browser window
Contents: Abstract from the article
Format: Flat text
Additional: should be searchable from the website

Fulltext (supplied if available)

Action: open available fulltext in a separate browser window
Contents: This could be a preprint, a postprint, or a publisher's version; if there are more, pick the most recent one added (unless manually blocked, e.g. private or institute only)
Format: usually pdf, perhaps other formats too? (doc ps ...)

Supplementary-materials (supplied if available)

Action: open child webpage in same browser window (i.e., a simple website-internal link)
Contents: This will be a regular plone webpage under end user control, but with additional queries to retrieve (link to) specific elements of the Pubman record at hand, e.g.

Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 31(3), 443-467. (possibly retrieved from PubMan, otherwise hardcoded)

The following supplementary materials are available for this paper:

- Dutch item sets for Experiments 1, 2, and 3 ([pdf](#))
- Two Dutch audio samples for Experiment 1 ([consistent](#) / [inconsistent](#))
- Cloze values of discourse-predictable nouns ([spss-sav](#))
- Additional data for all 120 items in Experiment 1 ([pdf](#))
- Additional data figure ([jpg](#))

Note: every one of the underlined strings (whose exact wording is free) must point to the appropriate pubman element. As there is no generic format here, end users will have to hyperlink these by hand. Also, note that the [supplementary-materials](#) link in the article citation should only be dynamically generated if the user has made the accompanying webpage; this requires some checking between pubman and website.

External-link (supplied if available)

Action: open the digital journal's webpage in a separate browser window, such that users with appropriate rights can access the journal fulltext, and others can pay to do so.

(1.2) An alternative website rendering, shown below,

- Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 31(3), 443-467. ([abstract](#) [fulltext](#) [external-link](#) [items](#) [consistent-audio-example](#) [inconsistent-audio-example](#) [cloze-data](#) [additional-data](#) [additional-figure](#))

is also possible, but seems less flexible. One drawback seems to be that, since the entry must be rendered on a webpage dynamically in response to a general PPI query (say, `get-citations(author = Jos van Berkum, status = {in-press, published}, keyword = JvB-personpage-selectivelisting)`), rendering it like this requires that all supplementary contents that the end user wants to list on the website (and no more) must be found automatically, and rendered with a custom label (e.g. "inconsistent-audio-example"). The construction via a standard supplementary-materials Plone webpage circumvents this problem. Second, the latter also adds textual flexibility (i.e., the user can tell a story about what the supplementary things are). I use the same flexibility in almost all below other genres. Note that the *archiving* of the various bits of supplementary materials still occurs in PubMan, and the stuff should have persistent identifiers there. However, we use a separate (and modifiable/eliminable) Plone webpage to render them coherently on the web.

(1.3) Article citations will need to be generated via a PPI query to pubman from a variety of website pages, e.g., a group page, a project page, a person page. Furthermore, end users will often want control over what is shown exactly. Therefore, the PPI query language should be such that the query can be customized to show, e.g.,

- only the articles for a particular group, or project, or person
- only the articles from a particular year
- only the articles with a specific user-definable keyword (e.g. `JvB-personpage-selectivelisting`, `JvB-myperceptionresearch`, `JvB-myproductionresearch`).

This allows people or groups or projects to retrieve selected subsets of citations based on their research themes or whatever else they want to highlight; a very important feature!

(1.4) Of course, beyond a sufficiently powerful PPI, this requires that (1) the end user should tag the relevant PubMan records with an appropriate keyword (e.g. `JvB-myperceptionresearch`), and (2) the end user should place the right PPI query somewhere on the relevant webpage. As the ways in which people, groups, or projects might partition their lists of articles (or other output) is not a closed set, all this implies that PubMan records should have a liberally extendible list of free keywords, accommodated via a "keywords" variable (with a flexible number of strings per value).

(1.5) A problem we need to deal with is that several MPI-PL researchers might co-author the same article, and that while author X might want to add keywords A, B and C for use on her personal page, author Y might want to add keywords D, E, and F. If we want just one PubMan record for this article (which makes sense), we need to know whom the keywords belong to (hence the prefix `JvB-`, there may be much more elegant solutions, I don't know).

(1.6) The above also raises an issue of ownership. If I add my publication with Peter Hagoort in PubMan, it seems rational that Peter need not do so too. Thus, a record may have multiple owners, i.e., MPI employees responsible for the record in PubMan.. Note that 'owners' may be different from 'creators', i.e., whereas a secretary or librarian may enter the record, others may own them.

(2) Book

(2.1) The typical website rendering is an APA-style reference/citation with some additional links:

- Levinson, S.C. (2003). *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge: CUP. ([abstract](#) [TOC](#) [supplementary-materials](#) [external-link](#))

[Abstract](#) (always supplied, hence obligatory content in PubMan)

Action: open abstract in a separate browser window
Contents: Abstract from the article
Format: Flat text
Additional: should be searchable from the website

[TOC](#) (always supplied, hence obligatory content in PubMan)

Action: open table of contents in a separate browser window
Contents: table of contents
Format: Flat text or seachable pdf
Additional: should be searchable from the website

[Supplementary-materials](#) (supplied if available)

(as (1))

[External-link](#) (supplied if available)

Action: open e.g., publisher's or bookseller's webpage in a separate browser window

(2.2) These citations would usually mingle with other citations (e.g. journal articles, chapters), most likely as the result of the same PPI query. See (1) for further details.

(2.3) Books can be edited or authored; this should be marked

(2.4) Additional question (also for other genres): can PubMan generate an APA-style citation, or should the PPI compile this out of bits and pieces? (author, year, etc).

(3) Chapter

(3.1) The typical website rendering is an APA-style reference/citation with some additional links:

- Van Berkum, J.J.A. (2004). *Sentence comprehension in a wider discourse: Can we use ERPs to keep track of things?* In M. Carreiras & C. Clifton Jr. (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERPs and beyond* (pp. 229-270). New York: Psychology Press.
([abstract](#) [fulltext](#) [supplementary-materials](#) [external-link](#))

[Abstract](#) (always supplied, hence obligatory content in PubMan)

(as (1))

[Fulltext](#) (supplied if available)

(as (1))

[Supplementary-materials](#) (supplied if available)

(as (1))

[External-link](#) (supplied if available)

Action: open e.g., publisher's or bookseller's webpage in a separate browser window

(4) Proceedings

(4.1) The typical website rendering is an APA-style reference/citation with some additional links:

- [Cutler, A., McQueen, J.M. & Zondervan, R. \(Eds.\) \(2000\). Proceedings of the SWAP Workshop, Nijmegen, The Nijmegen, The Netherlands. Nijmegen: Max Planck Institute for Psycholinguistics \(abstract TOC supplementary-materials external-link\)](#)

[Abstract](#) (always supplied, hence obligatory content in PubMan)

Action: open abstract in a separate browser window
Contents: Abstract from the article
Format: Flat text
Additional: should be searchable from the website

[TOC](#) (always supplied, hence obligatory content in PubMan)

Action: open table of contents in a separate browser window
Contents: table of contents
Format: Flat text or searchable pdf
Additional: should be searchable from the website

[Supplementary-materials](#) (supplied if available)

(as (1))

[External-link](#) (supplied if available)

Action: opens, e.g., conference/workshop website in separate browser window

(4.2) Contentwise, this is essentially the same as an edited book

(5) Proceedings paper

(5.1) The typical website rendering is an APA-style reference/citation with some additional links:

- McQueen, J.M., Cutler, A. & Norris, D. (2000). Positive and negative influences of the lexicon on phonemic decision-making. In Proceedings of ICSLP2000 (International Conference on Spoken Language Processing), Vol. 3, pp. 778--781, Beijing, China. ([abstract](#) [fulltext](#) [supplementary-materials](#) [external-link](#))

[Abstract](#) (always supplied, hence obligatory content in PubMan)

(as (1))

[Fulltext](#) (supplied if available)

(as (1))

[Supplementary-materials](#) (supplied if available)

(as (1))

[External-link](#) (supplied if available)

Action: opens, e.g., conference/workshop website in separate browser window

(6) Talk

(6.1) The typical website rendering is an APA-style reference/citation with some additional links:

- Van Berkum, J.J.A. (2007, invited). *Beyond Decontextualized Sentences: What can ERPs Tell Us about Pragmatics (and Semantics)?* Experimental Pragmatics 2007 (XPRAG-2007), Berlin, December 13-16. ([abstract](#) [slides](#) [handout](#) [supplementary-materials](#) [external-link](#))

Abstract (always supplied, hence obligatory content in PubMan)

(as (1))

Slides (supplied if available)

Action: open available slides in a separate browser window

Contents: Slides of the talk

Format: usually pdf or ppt (powerpoint), possibly Apple keynote

Handout (supplied if available)

Action: open available handout in a separate browser window

Contents: Handout for the talk

Format: usually pdf or doc

Supplementary-materials (supplied if available)

(as (1))

External-link (supplied if available)

Action: opens, e.g., conference/workshop website in separate browser window

(7) Poster

(7.1) The typical website rendering is an APA-style reference/citation with some additional links:

- Van Berkum, J.J.A., Holleman, B.C., Murre, J.M.J., Nieuwland, M.S., & Otten, M. (2007). *So how do you feel about this? An ERP study on opinion poll comprehension*. Annual Meeting of the Cognitive Neuroscience Society (CNS-2007), New York, May 5-9. ([abstract](#) [poster](#) [supplementary-materials](#) [external-link](#))

Abstract (always supplied, hence obligatory content in PubMan)
(as (1))

Poster (supplied if available)

Action: open available poster in a separate browser window
Contents: The poster itself
Format: pdf, ppt (powerpoint), Illustrator, etc...

Supplementary-materials (supplied if available)
(as (1))

External-link (supplied if available)

Action: opens, e.g., conference/workshop website in separate browser window

(8) Teaching

(8.1) The typical website rendering is roughly like the 2006 Annual Report entry (p. 185-187), but with optional extensions (to add slides, handouts, etc; none of this is obligatory):

- [Bowerman, M. \(2007\) Language acquisition and semantic typology. Masters' program in Cognitive Neuroscience, Radboud U. Nijmegen. \(description slides handout supplementary-materials external-link\)](#)

Description (supplied if available)

Action: open description in a separate browser window
Contents: Short description about the item
Format: Flat text
Additional: should be searchable from the website

Slides (supplied if available)

Action: open available slides in a separate browser window
Contents: Slides of the lecture (if it is a single lecture; if, not, see below)
Format: usually pdf or ppt (powerpoint), possibly Apple keynote

Handout (supplied if available)

Action: open available handout in a separate browser window
Contents: Handout for the lecture (if it is a single lecture; if, not, see below)
Format: usually pdf or doc

Supplementary-materials (supplied if available)

Action: open child webpage in same browser window (i.e., a simple website-internal link)
Contents: This will be a regular plone webpage under end user control, but with additional queries to retrieve (link to) specific elements of the Pubman record at hand, e.g.

[Bowerman, M. \(2007\) Language acquisition and semantic typology. Masters' program in Cognitive Neuroscience, Radboud U. Nijmegen.](#) (possibly retrieved from PubMan, otherwise hardcoded)

Here are the slides and handouts for this course:

- Lecture 1 ([slides handout](#))
 - Lecture 2 ([slides handout](#))
 - Lecture 3 ([slides handout](#))
-

Note: every one of the underlined strings (whose label is free) must translate to an underlying hyperlink to the appropriate pubman element. As there is no generic format here, end users will have to hyperlink these by hand. Also, note that the [supplementary-materials](#) link in the teaching citation should only be dynamically generated if the user has made the accompanying webpage; this requires some checking between pubman and website.

External-link (supplied if available)

Action: opens, e.g., electronic study guide in separate browser window

(9) Thesis

(9.1) The typical website rendering is an APA-style reference/citation with some additional links, as part of a list of many more such citations (and identical to the Book entry):

- [Van Berkum, J.J.A. \(1996\). *The psycholinguistics of grammatical gender: Studies in language comprehension and production*. Doctoral Dissertation, Max Planck Institute for Psycholinguistics. Nijmegen, The Netherlands: Nijmegen University Press. \(\[abstract\]\(#\) \[TOC\]\(#\) \[supplementary-materials\]\(#\) \[external-link\]\(#\)\)](#)

[Abstract](#) (always supplied, hence obligatory content in PubMan)

Action: open abstract in a separate browser window
Contents: Abstract from the article
Format: Flat text
Additional: should be searchable from the website

[TOC](#) (always supplied, hence obligatory content in PubMan)

Action: open table of contents in a separate browser window
Contents: table of contents
Format: Flat text or seachable pdf
Additional: should be searchable from the website

[Supplementary-materials](#) (supplied if available)

(as (1))

[External-link](#) (supplied if available)

Action: open e.g., publisher's or bookseller's webpage in a separate browser window

(10) Working paper

(10.1) The typical website rendering is an APA-style reference/citation (check the below) with some additional links:

- Van Berkum, J.J.A. (2008). *PubMan content categories needed for the MPIPL website and initial specs for website rendering of Pubman content*. Unpublished working paper, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. ([abstract](#) [fulltext](#) [supplementary-materials](#) [external-link](#))

[Abstract](#) (always supplied, hence obligatory content in PubMan)
(as (1))

[Fulltext](#) (supplied if available)
(as (1))

[Supplementary-materials](#) (supplied if available)
(as (1))

[External-link](#) (supplied if available)

Action: open e.g., publisher's or bookseller's webpage in a separate browser window

(11) Report

(11.1) The typical website rendering is an APA-style reference/citation (check the below) with some additional links:

- Van Berkum, J.J.A. (2006). *Sense and reference in language comprehension: A neurocognitive research program*. Final Vidi-grant report for Dutch Science Foundation. ([abstract](#) [fulltext](#) [supplementary-materials](#) [external-link](#))

[Abstract](#) (supplied if available)
(as (1))

[Fulltext](#) (supplied if available)
(as (1))

[Supplementary-materials](#) (supplied if available)
(as (1))

[External-link](#) (supplied if available)
Action: open e.g., project website

(12) Journal editing

(12.1) This is one of the additional science activities tracked by the institute (and possibly listed on the website):

- Van Berkum, J.J.A. (2006-). *Journal of Memory and Language*. Editorial board member. ([external-link](#))

[External-link](#) (supplied if available)

Action: open journal website

(12.2) As part of annual report compilation, this entry is usually reported on every year (e.g., "I did this in 2006", "I did this in 2007", etc). Presumably we don't want separate records for that. Should the existing entry be modified when appropriate? The same problem occurs with series editing. The most easy solution (also from a retrieval query perspective) seems to be to add new years every year, so "2004", "2005", 2008".

(13) Special issue

(13.1) The typical website rendering is an APA-style reference/citation (check the below) with some additional links:

- McQueen, J.M. & Cutler, A. (Eds.) (2001). Spoken Word Access Processes. Special issue of the journal *Language and Cognitive Processes*. Hove: Psychology Press. ([abstract](#) [TOC](#) [supplementary-materials](#) [external-link](#))

[Abstract](#) (always supplied, hence obligatory content in PubMan)

Action: open abstract in a separate browser window

Contents: Abstract from the article

Format: Flat text

Additional: should be searchable from the website

[TOC](#) (always supplied, hence obligatory content in PubMan)

Action: open table of contents in a separate browser window

Contents: table of contents

Format: Flat text or searchable pdf

Additional: should be searchable from the website

[Supplementary-materials](#) (supplied if available)

(as (1))

[External-link](#) (supplied if available)

Action: open e.g., the journal's webpage in a separate browser window

(13.2) Contentwise, this is essentially the same as an edited book

(14) Series editing

(14.1) This is one of the additional science activities tracked by the institute (and possibly listed on the website):

- [Kastens, K. \(2004-\). *Nijmegen working papers in information retrieval*. \(description supplementary-materials external-link\)](#)

[Description](#) (supplied if available)

Action: open description in a separate browser window

Contents: Short description of the series

Format: Flat text

Additional: should be searchable from the website

[Supplementary-materials](#) (supplied if available)

(as (1))

[External-link](#) (supplied if available)

Action: open e.g., the series' webpage in a separate browser window

(15) Event organized

(15.1) This is one of the additional science activities tracked by the institute (and possibly listed on the website):

- [Kastens, K. \(2008\). *First international workshop on PubMan-based web information retrieval*. Nijmegen, The Netherlands, March 11-13. \(\[description\]\(#\) \[supplementary-materials\]\(#\) \[external-link\]\(#\)\)](#)

[Description](#) (always supplied, hence obligatory content in PubMan)

Action: open description in a separate browser window

Contents: Short description of the series

Format: Flat text

Additional: should be searchable from the website

[Supplementary-materials](#) (supplied if available)

(as (1))

[External-link](#) (supplied if available)

Action: open e.g., the relevant event webpage in a separate browser window

(16) Honor/award

(16.1) This is one of the additional science activities tracked by the institute (and possibly listed on the website):

- [Laura Herbst \(2006\). *Best poster presentation at the 1st Max Planck PhD net workshop, Köln.* \(description supplementary-materials external-link\)](#)

[Description](#) (always supplied, hence obligatory content in PubMan)

Action: open description in a separate browser window
Contents: Short description of the honor/award (e.g. see annual report 2006, p. 159)
Format: Flat text
Additional: should be searchable from the website

[Supplementary-materials](#) (supplied if available)

(as (1)). Possibly also links to media coverage, jury report, etc., attached to this record in PubMan

[External-link](#) (supplied if available)

Action: open e.g., the grant agency's relevant webpage in a separate browser window

(17) Media coverage

(17.1) This is one of the additional science activities tracked by the institute (and possibly listed on the website):

- Hagoort, P. (2008). *Het brein ontleed*. Volkskrant interview, March 11. ([media-coverage supplementary-materials external-link](#))

[Media-coverage](#) (supplied if available)

Action: open available 'fulltext' materials in a separate browser window

Contents: This could be a scan of the interview, a video recording, a radio interview, etc.

Format: pdf, jpg, mpg, avi, mp3, doc, etc. etc.

[Supplementary-materials](#) (supplied if available)

(as (1)). A flexible way to add more stuff (e.g., if there is a radio interview plus a transcript)

[External-link](#) (supplied if available)

Action: open e.g., the original source (e.g., a magazine archive) in a separate browser window

(18) Software/corpus

(18.1) The typical website rendering is an APA-style reference/citation (check the below) with some additional links:

- Petersson, K.M. (2008). *The pipeline: A MatLab library for neuroimaging experiments*. ([description](#) [download](#) [supplementary-materials](#) [external-link](#))

[Description](#) (supplied if available)

(as (1))

[Download](#) (supplied if available)

Action: allows for downloading a file or archive

Contents: a (set of) program(s) or a corpus

Format: could be anything

[Supplementary-materials](#) (supplied if available)

(as (1)) Adds more flexibility.

[External-link](#) (supplied if available)

Action: open e.g., a relevant tools webpage for downloading

Ordering, additional tags, searchability

Here's a possible ordering/grouping for our interface, in two bins, and then alphabetical

--- research products ---

Book (edited or authored)

Chapter

Journal article

Poster

Proceedings (edited)

Proceedings paper

Report

Special issue (edited)

Talk

Thesis

Working paper

--- other items ---

Events organized

Honors/awards

Journal editing

Media coverage

Series editing

Software

Teaching

All primary research products genres (the top part of the list) have an 'abstract' field, all other items have a 'description' field instead.

Tags: As for tags, note that a single record can be associated with more than one MPI person, MPI group, and MPI project.

Also, accessibility needs to be at least three-level: private / institute / world
Only records or parts thereof tagged with 'world' make it to the website.

Search: Plone website users will want to search the site, which presents a problem when part of the site is dynamically generated from PubMan. We suggest that the website search machine has access to:

- PubMan-stored citations (essentially the orange bits at the top of each page)
 - PubMan-stored abstracts/descriptions
 - PubMan-stored TOC's
-