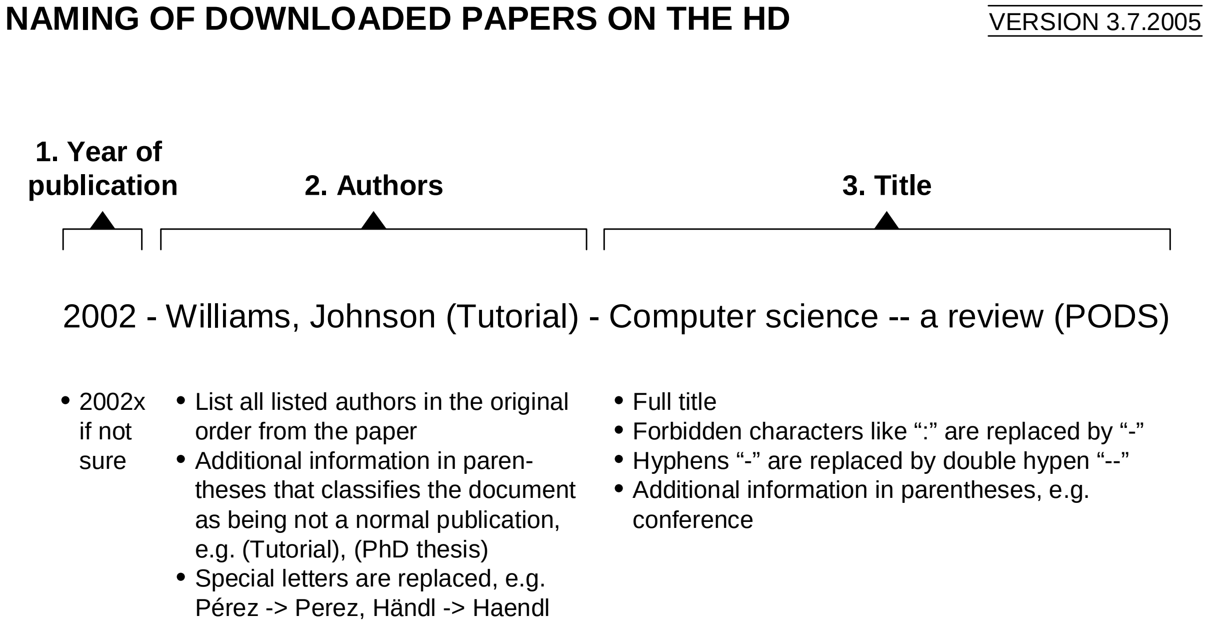
How to name downloaded papers on your HD

First version: December 11, 2004
This version: July 3, 2005
Wolfgang Gatterbauer

Almost every researcher saves downloaded papers on his HD in a different and incoherent way. Here are (1) a recommendation for a naming scheme, (2) a rough outline of ways to organize downloaded papers, and (3) the reasons for the choice.

1. RECOMMENDATION

All downloaded papers should be simply dumped into one single folder, but the names should be changed to contain 1. Year of publication, 2. Authors and 3. Title


NAMING OF DOWNLOADED PAPERS ON THE HD

VERSION 3.7.2005

2002 - Williams, Johnson (Tutorial) - Computer science -- a review (PODS)

- 2002x if not sure

- List all listed authors in the original order from the paper
- Additional information in parentheses that classifies the document as being not a normal publication, e.g. (Tutorial), (PhD thesis)
- Special letters are replaced, e.g. Pérez -> Perez, Händl -> Haendl

- Full title
- Forbidden characters like “:” are replaced by “-”
- Hyphens “-” are replaced by double hypen “--”
- Additional information in parentheses, e.g. conference

Other made up example file names that follow the proposed naming scheme:

```
2001 - Jackson - Very simple title
2002 - Johnson (PhD Thesis) - Model for doing something
2002x- Williams, Lambert - VCS -- the Very-Clever-System
2003 - Adams, Newton - Tools for doing one, two and three
2003 - Tintifax, Kasperl (Tutorial) - Knowledge Roadmap (PODS)
```

2. THINGS TO CONSIDER

General aspects about organizing downloaded papers

- Naming schemes
  - Keeping the unchanged, original name
  - Renaming papers when saving
    - changing the name coherently vs. in an ad-hoc fashion
- Folder structure
  - Having different folders vs. keeping all papers in one folder
  - Having some well-conceived-of order system (MECE as much as possible) vs. creating these folders on an ad-hoc basis
- Using some software like *EndNote* to automatically organize papers on HD (?)
- Having local copies of papers vs. not downloading papers at all

3. RATIONALE FOR CHOICE

Rationale for preferring a consistent naming scheme over using folders over using folders for categorizing papers into topics

- Essence:
  - **The advantages of having all papers consistently named in one folder outweigh the upfront investment of ~15 sec for consistently renaming a paper when saving it to the HD**
- Advantages
  - Text-based search is very handy
    - "Ctrl" + "F" can be used in Windows Explorer or locate in Unix
    - Outlook plug-in *Lookout* fully pre-indexes the search for file names
  - No inconsistencies in folder structure possible
    - Some papers would always have appear in different folders (e.g. in folder "Tabular IE" and "NLP IE") / Finding a MECE folder structure not possible because research subjects consistently changing
  - Improved collaboration
    - Whenever several people have to deal with the same paper it is easier if a common naming scheme is adhered to as the name is self-explanatory (e.g. eliminating duplicates when consolidating files from different people)
    - Local systems that organize papers in folders lose their value when people exchange files
  - Using folders can be an additional option for non-content categories, e.g. distinction between read and unread papers
- Disadvantages
  - Time to change name when saving to HD
  - It's perhaps more difficult to find papers belonging to the same topic
    - *EndNote* might help (?)
    - *TWiki* permits full content indexing -> Full content search -> Resulting file names immediately give idea about content (similar to Google results with content excerpt)
  - Some systems might not be able to handle long names (? Unix: space, Wiki: special characters)

Proposed building blocks of a name; necessary information for easy re-discovery

- 1. **Year** of Publication
  - Permits to quickly grasp the novelty of a paper and skip older ones
  - Only 7 digits in the beginning, e.g. "2004 - ". Stay with 7 digits even if using "x" to mark that you are not sure about the date (instead of "?"), e.g. "2003x- "
- 2. **Authors**
  - Surnames of all authors
    - so you can easily search for all papers from a specific author
    - separated by commas
  - No first names despite possible ambiguities (reasonable trade-off to save space)
  - To search for special authors one uses the search function. Authors of interest might not be always the first author anyway.
- 3. **Title**
  - Complete title
    - so you can make an easy text search for topic words
  - Forbidden signs like ":" replaced with "--", e.g. "VCS -- the Very-Clever-System"
- Additional, optional information in parentheses
  - After 2. Authors
    - information that classifies the document as not being a normal publication, e.g. (PhD Thesis), (Tutorial)
  - After 3. Title
    - conference information, e.g. (PODS 2004)

Chosen order of building blocks

- Most reasonable order (shortlist):
  - A: **Year – Authors – Title**
  - B: Authors – Year – Title
- Reasons for option A (Year as first block) instead of option B (Authors as first block)
  - Advantages
    - Authors might as well be second authors -> one anyway has to use a textual search function (e.g. "ctrl" + "F") when searching for papers from certain authors
    - Authors still remain readable in a very fast manner, because the Year block is exactly 7 characters in the beginning of the name ("2003 - " still looks like column; in contrast to the difficulty when looking for specific years if Authors are first block)
  - Disadvantages
    - Date is not the most important characteristic of a paper and takes away 7 characters in the beginning of the name; might be a problem in circumstances with readability of only a limited number of characters (e.g. documents in a directory on a web server)