From: Konstantaki, Laura Amalia
Sent: Dienstag, 17. Oktober 2023 06:14
To: Last, Sarah; Hösli, Frank
Cc: Felder, Fabian; Cantini, Federico
Subject: RE: A task for Frank

Good morning together,

Frank I am sorry if the redmine ticket was not updated, I thought since the information in the email was specific this should be clear (and in the redmine ticket are lilnks that I need to check at some later point anyway).

As Sarah writes and also as I wrote in my email (on 29.9) please do only the following (here are publications only with a PDF – there is no discussion about no full text publications- and from these check only the ones with a DOI):

1. Replace automatically in DORA all PDFS possible for the publications having a DOI with the following publishers: Elsevier, Springer Nature, American Chemical Society, Taylor & Francis, Copernicus, Springer. Would be this link (but only for the ones having a DOI, since for these ones the automatic download will work). Here as Sarah says, for the ones a PDF is exchanged, please also remove the tag #check_pdf.

https://www.dora.lib4ri.ch/islandora/search?islandora_solr_search_navigation=1&f[0]=mods_originInfo_encoding_w3cdtf_keyDate_yes_dateIssued_dt%3A[2006-01-01T00%3A00%3A00Z%20TO%202012-01-01T00%3A00%3A00Z]&f[1]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Apublications%22&f[2]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Aexternal%22&f[3]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Apublications%22&f[4]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-int%22&f[5]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-ext%22&f[6]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/eawag\%3Aext%22&f[7]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Aforum%22&f[8]=mods_relatedItem_host_originInfo_publisher_s%3A(%22Elsevier%22OR%22Springer%20Nature%22OR%22American%20Chemical%20Society%22OR%22Taylor\%20%26\%20Francis%22OR%22Copernicus%22OR%22Springer%22)

2. (a) Download automatically all PDFS possible for the publications having a DOI with the following publishers: Wiley, IOP publishing AIP, SPIE, RSC; (b) Automatically delete the first page if it is a cover page -> I remember that you had such a code for the migration of PSI, or not? We can use the same code we used then with the same parameters; (c) replace the PDFs in DORA & remove the tag #check_pdf . Here again there is no discussion about no full text PDFs. Would be this link (but only for the ones having a DOI, since for these ones the automatic download will work):

https://www.dora.lib4ri.ch/islandora/search?islandora_solr_search_navigation=1&f[0]=mods_originInfo_encoding_w3cdtf_keyDate_yes_dateIssued_dt%3A[2006-01-01T00%3A00%3A00Z%20TO%202012-01-01T00%3A00%3A00Z]&f[1]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Apublications%22&f[2]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Aexternal%22&f[3]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Apublications%22&f[4]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-int%22&f[5]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-ext%22&f[6]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/eawag\%3Aext%22&f[7]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Aforum%22&f[8]=mods_relatedItem_host_originInfo_publisher_s%3A(%22Wiley%22OR%22IOP%20Publishing%22OR%22American%20Institute%20of%20Physics%22OR%22SPIE%22OR%22Royal%20Society%20of%20Chemistry%22)

Please let me know if this is not clear. I will update the redmine ticket later today to be in sync.

LG
Laura

////
Dr. Laura Konstantaki · Group Leader Publication Services
Lib4RI - Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

Lib4RI: Eawag-Empa · Überlandstrasse 133 · 8600 Dübendorf · Switzerland
+41 58 765 55 90 · laura.konstantaki@lib4ri.ch · www.lib4ri.ch

Working days: Tuesday, Thursday, Friday

From: Last, Sarah <Sarah.Last@lib4ri.ch>
Sent: Montag, 16. Oktober 2023 07:53
To: Hösli, Frank <Frank.Hoesli@lib4ri.ch>; Konstantaki, Laura Amalia <Laura.Konstantaki@lib4ri.ch>
Cc: Felder, Fabian <Fabian.Felder@lib4ri.ch>; Cantini, Federico <Federico.Cantini@lib4ri.ch>
Subject: RE: A task for Frank

Hi Frank

Thank you very much!
As far as I understood, only the two links Laura sent need to be checked and PDF replaced.

https://www.dora.lib4ri.ch/islandora/search?islandora_solr_search_navigation=1&f%5b0%5d=mods_originInfo_encoding_w3cdtf_keyDate_yes_dateIssued_dt%3A%5b2006-01-01T00%3A00%3A00Z%20TO%202012-01-01T00%3A00%3A00Z%5d&f%5b1%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Apublications%22&f%5b2%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Aexternal%22&f%5b3%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Apublications%22&f%5b4%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-int%22&f%5b5%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-ext%22&f%5b6%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/eawag\%3Aext%22&f%5b7%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Aforum%22&f%5b8%5d=mods_relatedItem_host_originInfo_publisher_s%3A(%22Wiley%22OR%22IOP%20Publishing%22OR%22American%20Institute%20of%20Physics%22OR%22SPIE%22OR%22Royal%20Society%20of%20Chemistry%22)

https://www.dora.lib4ri.ch/islandora/search?islandora_solr_search_navigation=1&f%5b0%5d=mods_originInfo_encoding_w3cdtf_keyDate_yes_dateIssued_dt%3A%5b2006-01-01T00%3A00%3A00Z%20TO%202012-01-01T00%3A00%3A00Z%5d&f%5b1%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Apublications%22&f%5b2%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Aexternal%22&f%5b3%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Apublications%22&f%5b4%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-int%22&f%5b5%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-ext%22&f%5b6%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/eawag\%3Aext%22&f%5b7%5d=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Aforum%22&f%5b8%5d=mod

s_relatedItem_host_originInfo_publisher_s%3A(%22Elsevier%22OR%22Springer%20Nature%22OR%22American%20Chemical%20Society%22OR%22Taylor\%20%26\%20Francis%22OR%22Copernicus%22OR%22Springer%22)

These two links do not contain no fulltexts. Regarding the match of the top link in Redmine 318 I think this is not the newest link anymore: http://lib-dora-dev1.emp-eaw.ch:3000/issues/318 (Laura, I am happy when you confirm this).
So there is no need to check the 223 publishers, it's enough when these two links mentioned here are all replaced and the tag is then removed (#check_pdf).

Kind regards
Sarah

From: Hösli, Frank <Frank.Hoesli@lib4ri.ch>
Sent: Freitag, 13. Oktober 2023 19:03
To: Konstantaki, Laura Amalia <Laura.Konstantaki@lib4ri.ch>; Last, Sarah <Sarah.Last@lib4ri.ch>
Cc: Felder, Fabian <Fabian.Felder@lib4ri.ch>; Cantini, Federico <Federico.Cantini@lib4ri.ch>
Subject: RE: A task for Frank

Hoi Laura

A short feedback. Based on the two links from Laura, this is the basic search: (rather to document it for me, I don't expect that you analyze it):

http://lib-dora-prod1.emp-eaw.ch:8080/solr/collection1/select?wt=csv&indent=true&csv.separator=|&sort=PID+asc&rows=987654321&q=PID:*%5c%3a*
+AND+mods_identifier_doi_mt:*
+AND+fedora_datastream_latest_FULL_TEXT_ID_mt:*
+AND+mods_originInfo_encoding_w3cdtf_keyDate_yes_dateIssued_dt:[2006-01-01T00%3A00%3A00Z%20TO%202012-01-01T00%3A00%3A00Z]
+NOT+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/psi%5c%3Apublications%22
+NOT+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/psi%5c%3Aexternal%22
+NOT+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Apublications%22
+NOT+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Awsl%5c-int%22
+NOT+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Awsl%5c-ext%22
+NOT+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/eawag%5c%3Aext%22
+NOT+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Aforum%22
&fl=PID
%2c+mods_identifier_doi_mt
%2c+RELS_EXT_fullText_literal_mt
%2c+mods_relatedItem_host_originInfo_publisher_mt
%2c+mods_originInfo_publisher_mt
%2c+mods_part_extent_start_mt
%2c+mods_part_extent_end_mt
%2c+mods_relatedItem_host_part_extent_start_mt
%2c+mods_relatedItem_host_part_extent_end_mt
%2c+RELS_EXT_isMemberOfCollection_uri_mt
%2c+mods_note_additional?information_mt

This should match the top link in Redmine #318 plus DOI requirement plus addtional full-text-fatastream-filter (as indication of a PDF, it seems actually there are a few publications without PDF).

Concerning the publisher filtering however, it is probably more flexible to do it code-wise afterwards (applying it on the current Solr results).

Some advice/confirmation would be appreciated actually because among the publications where to replace the PDF possibly there is a large varity of publisher names and how they are written/abbreviated - I counted 223.
In link **1.** and **2.** below there are only 11 mentioned. How to deal with the others? Or do we stick with these 11 publishers and only try to download+replace PDF from them?

Kind Regards,
Frank

```
Publisher Names found: 223
Array(
    [1] => Academic Press
    [2] => ACS
    [3] => AIAA
    [4] => AIDIC
    [5] => AIP Publishing
    [6] => AMA Service GmbH
    [7] => American Association for the Advancement of Science
    [8] => American Chemical Society
    [9] => American Concrete Institute
    [10] => American Geophysical Union
    [11] => American Institute of Aeronautics and Astronautics
    [12] => American Institute of Physics
    [13] => American Meteorological Society
    [14] => American Physical Society
    [15] => American Physiological Society
    [16] => American Scientific Publishers
    [17] => American Society for Microbiology
    [18] => American Society for Testing and Materials
    [19] => American Society of Agricultural and Biological Engineers
    [20] => American Society of Civil Engineers ASCE
    [21] => American Society of Ichthyologists and Herpetologists
    [22] => American Society of Mechanical Engineers ASME
    [23] => American Society of Parasitologists
    [24] => American Society of Tropical Medicine and Hygiene
    [25] => American Water Works Association
    [26] => Annual Reviews
    [27] => AO Foundation Davos
    [28] => ASM International
    [29] => ASME
    [30] => Association Générale des Hygiénistes et Techniciens Municipaux
    [31] => Association pour la Diffusion de la Recherche Alpine
    [32] => Associação Brasileira de Cerâmica
    [33] => Associação Brasileira de Saúde Coletiva
    [34] => ASTM International
    [35] => Austrian Academy of Sciences Press
    [36] => Beilstein-Institut
    [37] => Bentham
    [38] => Blackwell
    [39] => Blackwell Publishing
    [40] => Blackwell Science Ltd.
    [41] => BMJ Publishing Group
    [42] => Boca Raton
    [43] => Brill
    [44] => Cambridge University Press
    [45] => Canadian Science Publishing
    [46] => Centre for Health and Population Research
    [47] => Chinese Society of Pavement Engineering
    [48] => Cold Spring Harbor Laboratory Press
    [49] => Columbia University in the City of New York
    [50] => Copernicus
    [51] => Corporation for National Research Initiatives
    [52] => CRC Press
    [53] => CRC Press; Woodhead Publishing
    [54] => CSIRO
```

[55] => Czech Geological Survey
[56] => de Gruyter
[57] => de Gruyter Recht
[58] => Desalination Publications
[59] => Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall DWA
[60] => Dordrecht
[61] => Ecological Society of Japan
[62] => Ecology and Civil Engineering Society
[63] => Edition Sigma
[64] => EDP Sciences
[65] => Electrochemical Society
[66] => Elsevier
[67] => Emerald
[68] => EMH Swiss Medical Publishers
[69] => Empa
[70] => Empa; BAFU
[71] => EMW Publishing
[72] => EPF Lausanne
[73] => Erich Schmidt Verlag
[74] => espazium Verlag
[75] => ETH Zurich
[76] => ETH Zürich
[77] => European Commission
[78] => European Council for Modelling and Simulation
[79] => Finnish Society for Science and Technology Studies
[80] => Frontiers Media
[81] => Future Science
[82] => Geological Society
[83] => Geological Society of America
[84] => Geological Society of London
[85] => German Medical Science
[86] => Global Nest
[87] => Hindawi
[88] => Hirzel Verlag
[89] => Hogrefe
[90] => Hohai University
[91] => Horizon House Publications Ltd
[92] => ICE Publishing
[93] => IEEE
[94] => IGI Global
[95] => IMR Press
[96] => Inderscience
[97] => Information Science Publishing
[98] => Institut Français du Pétrole IFP
[99] => Institute of Electrical and Electronics Engineers, Inc.
[100] => Institute of Electrical Engineers of Japan
[101] => Institute of Noise Control Engineering
[102] => Institute of Physics Publishing
[103] => Institution of Engineering and Technology (IET)
[104] => InTech
[105] => IntechOpen
[106] => Inter-Research Science Publishing
[107] => International association for bridge and structural engineering (IABSE)
[108] => International Astronomical Union
[109] => International Centre for Applied Thermodynamics
[110] => International federation of automatic control (IFAC)
[111] => International Glaciological Society
[112] => International Society for Horticultural Science
[113] => International Society of Histology and Cytology
[114] => IOP
[115] => IOP Publishing
[116] => Iron and Steel Institute of Japan
[117] => ISTE Ltd.
[118] => IWA Publishing
[119] => Japan Institute of Metals
[120] => Japan Petroleum Institute
[121] => Japan Society of Applied Physics; IEEE Electron Device Society
[122] => Japan Society of Mechanical Engineers
[123] => John Wiley & Sons
[124] => John Wiley & Sons Ltd

[125] => John Wiley and Sons
[126] => Karger
[127] => Landwirtschaftsverlag
[128] => Libertas Academica
[129] => Lippincott Williams & Wilkins
[130] => Macmillan Publishers; World Scientific
[131] => Madagascar Wildlife Conservation
[132] => Mary Ann Liebert
[133] => Materials Research Society
[134] => Mathematical Sciences Publishers (MSP)
[135] => MDPI
[136] => Microbiology Society
[137] => Mineralogical Society of America
[138] => National Academy of Sciences, USA
[139] => National Institute of Environmental Health Sciences
[140] => National Speleological Society
[141] => Naturforschende Gesellschaft in Bern
[142] => Nippon Kinzoku Gakkai
[143] => North American Benthological Society
[144] => NTNU
[145] => oekom Verlag
[146] => Optica Publishing Group
[147] => OSA
[148] => OSA/DH/FTS/HISE/NTM/OTA
[149] => OSA/OSF
[150] => Oxford University Press
[151] => PagePress
[152] => Pan Stanford
[153] => Pensoft
[154] => Polish Academy of Sciences
[155] => Portland Press
[156] => Practical Action Publishing
[157] => Princeton University Press
[158] => Public Library of Science
[159] => Regional Euro-Asian Biological Invasions Centre
[160] => Resilience Alliance
[161] => RILEM publications SARL
[162] => Routledge
[163] => Royal Society
[164] => Royal Society of Chemistry
[165] => RSC
[166] => SAE International
[167] => Sage
[168] => Saxe-Coburg Publications
[169] => Schweizerbart
[170] => Schweizerische Chemische Gesellschaft
[171] => Schweizerische Geologische Gesellschaft
[172] => Schweizerische Vereinigung Von Petroleum-Geologen und-Ingenieuren
[173] => Schweizerischer Verein für Vermessung und Kulturtechnik
[174] => Science Press
[175] => Scientific Research Publishing
[176] => SEPM
[177] => SGEM World Science
[178] => Shaker
[179] => Society ALTEX Edition
[180] => Society for Imaging Science and Technology
[181] => Society for Industrial and Applied Mathematics
[182] => Society of Exploration Geophysicists
[183] => Société de Physique et d'Histoire Naturelle de Genève
[184] => Société Francophone de Santé et Environnement
[185] => Société Française d'Ichthyologie
[186] => SPIE
[187] => SPIE - International Society for Optical Engineering
[188] => Springer
[189] => Springer Nature
[190] => Taiwan Association for Aerosol Research
[191] => Taylor & Francis
[192] => Technical University of Denmark
[193] => Techno-Press
[194] => Texas A & M University

```
    [195] => Thai Society of Higher Education Institutes on Environment
    [196] => The Electrochemical Society
    [197] => The Electrochemical Society (ECS)
    [198] => The International Bank for Reconstruction and Development / The World Bank
    [199] => The Japan society of mechanical engineers
    [200] => Thieme
    [201] => Trans Tech Publications
   [202] => Tsinghua University Press; Springer
    [203] => Unified Theory of Information Research Group
    [204] => University of Chicago Press
    [205] => University of Toronto Press
    [206] => Universität Basel
    [207] => vdf Hochschulverlag AG an der ETH Zürich
    [208] => Vilnius Gediminas Technical University
    [209] => VS Verlag
    [210] => Water Environment Federation
    [211] => Wiley
    [212] => Wiley-Blackwell
    [213] => Wiley-Interscience
    [214] => Wiley-VCH
    [215] => William Andrew
    [216] => WIT
    [217] => WIT Press
    [218] => Woodhead Publishing
    [219] => Woodhead Publishing Limited
    [220] => Woodhead Publishing; CRC Press
    [221] => World Health Organization
    [222] => World Scientific
   [223] => World Scientific Publishing
)
```

From: Konstantaki, Laura Amalia <Laura.Konstantaki@lib4ri.ch>
Sent: Freitag, 29. September 2023 09:17
To: Hösli, Frank <Frank.Hoesli@lib4ri.ch>; Last, Sarah <Sarah.Last@lib4ri.ch>
Cc: Felder, Fabian <Fabian.Felder@lib4ri.ch>; Cantini, Federico <Federico.Cantini@lib4ri.ch>
Subject: RE: A task for Frank

Hi Sarah and Frank,

Thanks a lot for the checks and ideas!

@Sarah, could you please save these checks/conclusions in our wiki -> maybe under diverses or if there is a place where we have info about full-text PDF. I find this information usefull to have also for the future.

@Frank, great thinking, nevertheless I would suggest to keep it “simple”, so please do the following:

1. Replace automatically in DORA all PDFS possible for the publications having a DOI with the following publishers: Elsevier, Springer Nature, American Chemical Society, Taylor & Francis, Copernicus, Springer. Would be this link (but only for the ones having a DOI, since for these ones the automatic download will work):

https://www.dora.lib4ri.ch/islandora/search?islandora_solr_search_navigation=1&f[0]=mods_originInfo_encoding_w3cdtf_keyDate_yes_dateIssued_dt%3A[2006-01-01T00%3A00%3A00Z%20TO%202012-01-01T00%3A00%3A00Z]&f[1]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Apublications%22&f[2]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Aexternal%22&f[3]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Apublications%22&f[4]=-

RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-int%22&f[5]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-ext%22&f[6]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/eawag\%3Aext%22&f[7]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Aforum%22&f[8]=mods_relatedItem_host_originInfo_publisher_s%3A(%22Elsevier%22OR%22Springer%20Nature%22OR%22American%20Chemical%20Society%22OR%22Taylor\%20%26\%20Francis%22OR%22Copernicus%22OR%22Springer%22)

2. (a) Download automatically all PDFS possible for the publications having a DOI with the following publishers: Wiley, IOP publishing AIP, SPIE, RSC; (b) Automatically delete the first page if it is a cover page -> I remember that you had such a code for the migration of PSI, or not? We can use the same code we used then with the same parameters; (c) replace the PDFs in DORA. Would be this link (but only for the ones having a DOI, since for these ones the automatic download will work):

https://www.dora.lib4ri.ch/islandora/search?islandora_solr_search_navigation=1&f[0]=mods_originInfo_encoding_w3cdtf_keyDate_yes_dateIssued_dt%3A[2006-01-01T00%3A00%3A00Z%20TO%202012-01-01T00%3A00%3A00Z]&f[1]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Apublications%22&f[2]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/psi\%3Aexternal%22&f[3]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Apublications%22&f[4]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-int%22&f[5]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Awsl\-ext%22&f[6]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/eawag\%3Aext%22&f[7]=-RELS_EXT_isMemberOfCollection_uri_ms%3A%22info\%3Afedora\/wsl\%3Aforum%22&f[8]=mods_relatedItem_host_originInfo_publisher_s%3A(%22Wiley%22OR%22IOP%20Publishing%22OR%22American%20Institute%20of%20Physics%22OR%22SPIE%22OR%22Royal%20Society%20of%20Chemistry%22)

3. After 1 and 2 are done, we will check again what is left from the original link and decide how to proceed.

LG
Laura

////
Dr. Laura Konstantaki · Group Leader Publication Services
Lib4RI - Library for the Research Institutes within the ETH Domain: Eawag, Empa, PSI & WSL

Lib4RI: Eawag-Empa · Überlandstrasse 133 · 8600 Dübendorf · Switzerland
+41 58 765 55 90 · laura.konstantaki@lib4ri.ch · www.lib4ri.ch

Working days: Tuesday, Thursday, Friday

From: Hösli, Frank <Frank.Hoesli@lib4ri.ch>
Sent: Donnerstag, 28. September 2023 19:08
To: Last, Sarah <Sarah.Last@lib4ri.ch>; Konstantaki, Laura Amalia <Laura.Konstantaki@lib4ri.ch>
Cc: Felder, Fabian <Fabian.Felder@lib4ri.ch>; Cantini, Federico <Federico.Cantini@lib4ri.ch>
Subject: RE: A task for Frank

Hello Together

Thanks for reminding me this PDF download project (..4 years ago).
Now it will (hopefully) be a little bit more ‘straight-forward’, but until we can take benefit of a coverpageGPT AI also now some human interactions will probably be required within the entire task.

A former remark stated that the size of page 1 (in bytes) is alone only a very vague indication for a cover page. The publisher and looking for a preset pattern of words/terms were better criterias.

For example https://pubs.rsc.org/en/content/articlelanding/2011/CC/c1cc12490k :
There is a 'Downloaded by' notation in a rather low amount of words, there is also the sequnece 'View Article Online', 'Journal Homepage', 'Table of Contents for this issue'. This is no certain but still a remarable indication of a cover page.

Over all we perhaps should proceed like this:

Step 1)

Downloading as many PDFs a possible.
Per PDF storing some metadata (DORA PID, publisher, DOI, ...) in file.

Step 2)

Cloning page 1 of each PDF and extracing its text
Manually setting up (better) patterns to identify a cover page.
Repeating this step until we have (per publisher probably) satisfing patterns where to cut off page 1.

Step 3)

Cutting off an assumed cover page and uploading each PDFs to its PID into DORA.

This approach is not final though, for example because...:

About 1)

There are 46 publications with a WoS ID only, and 329 with a Scopus ID only - how/when to deal with them?
Each PDF may have metadata in its file header - drop it or maintain it?

About 2)

- Re-tuning patterns may become more and more painful, so a 'page-1-keep/drop' special-case-list with PIDs may help with tricky cases.
- Generally this step will be quite work-intensive, so big page-1-thumbnails can be useful (not to rely on the lame(?) Windows-PDF-preview).
- What is the chance that a PDF has two or multiple cover pages?

About 3)

- In theory all PDF-derivatives could be pre-produced for less system-stress – but not sure if DORA will like 'fremd-ingested' derivatives.
- If the cover page detection in step 2) has expected/known lacks, we need a last recheck/postfix here.

Kind Regards,
Frank

P.S.: Below the cover-page identifying patters (publisher-based) used 4 years ago – not sure how far still reliable meanwhile:

```
$identAry['CellPress']   = array( "Article", "Graphical Abstract", "Authors", "Correspondence", "In Brief", "Highlights" /* , "Accession Numbers" */  );
```

```
$identAry['PopularPhysics']  = array( "To cite this article:", "View the article online for updates and enhancements.", "This
content was downloaded from IP address" );        // "Related content"
// ^^^^ also for Physica Scripta + 'epl' + 'Physics in Medicine & Biology'
$identAry['RevSciInst']               = array( "Cite as:", "Submitted:", "Accepted:", "Published Online:", "~ARTICLES YOU MAY BE
INTERESTED IN" );
$identAry['JourAppPhys']     = array( "Cite as:", "Published Online:", "~ARTICLES YOU MAY BE INTERESTED IN" );                //
Cite as: Journal of Applied Physics
$identAry['JourAppPhys2']    = array( "Cite as: Journal of Applied Physics", "Published Online:", "Journal of Applied Physics",
"American Institute of Physics" );
$identAry['ETH-Biblio']               = array( "Autor", "Objekttyp:", "PDF erstellt am:", "Persistenter Link:", "Ein Dienst der
ETH-Bibliothek" );
$identAry['WileyVCH-SI']     = array( "Supporting Information", "Copyright Wiley-VCH Verlag GmbH", "Co. KGaA, 69451 Weinheim, " );
// $identAry['WileyVCH-CPC']  = array( "DOI:", "Wiley-VCH Verlag GmbH", "Co. KGaA, Weinheim", "ChemPhysChem" );
$identAry['WileyGeneral']    = array( "![a]", "![*]", "!a[*]", "!a,b", "!a,*", "!b,*", "!**", "!1. Introduction", "Wiley-VCH
Verlag GmbH", "Co. KGaA", "Weinheim" );          // keep after WileyVCHsi!
$identAry['SPIEDigLib']               = array( "!1 Introduction", "Downloaded From:
https://www.spiedigitallibrary.org/journals/", "Terms of Use: https://www.spiedigitallibrary.org/terms-of-use" );
$identAry['ResearchGate']    = array( "See discussions, stats, and author profiles for this publication at",
"www.researchgate.net", "~SEE PROFILE" );
$identAry['JSTOR']           = array( "Your use of the JSTOR archive indicates your acceptance of the Terms", "This content
downloaded from" );
$identAry['Forstverein']     = array( "www.forstverein.ch", "Wissenschaftliche Zeitschrift des Schweizerischen Forstvereins",
"Forstwesen abonnieren" );
$identAry['BioOne']          = array( "BioOne sees sustainable scholarly publishing as an inherently collaborative enterprise",
"Downloaded From:", "bioone.org/terms-of-use" );
$identAry['AmInstPhys']      = array( "Citation:", "View online:", "View Table of Contents:", "Published by the American Institute
of Physics" );              // "Articles you may be interested in"
$identAry['AmVaccuSoc']      = array( "Citation:", "View online:", "View Table of Contents:", "Published by the American Vacuum
Society" );                   // "Articles you may be interested in"
$identAry['AddBarLow']       = array( "Citation:", "View online:", "View Table of Contents:", "Published by the", "~Articles you
may be interested in" );
$identAry['TaylorFancis']    = array( "cite this article:", "link to this article:", "Published online:", "Submit your article to
this journal", "Article views:" );              // "Citing articles:"
$identAry['www.RSC.org']     = array( "Volume", "Number", "Pages", "www.rsc.org/" );
$identAry['www.RSC.org2']    = array( "www.rsc.org/", "Volume", "Number", "Pages" );
$identAry['SoftMatter']               = array( "View Article Online", "Volume", "Number", "Page", "www.softmatter.org" );
$identAry['SoftMatter2']     = array( "View Article Online", "www.softmatter.org", "Volume", "Number", "Page" );
$identAry['PhysChChPhys']    = array( "View Article Online", "Table of Contents for this issue", "is published as part of a",
"PCCP" );             // space may miss after 'a'
$identAry['DaltonTrans']     = array( "View Article Online", "Table of Contents for this issue", "is published as part of the
Dalton Trans" );
$identAry['DaltonTrans2']    = array( "View Article Online", "Table of Contents for this issue", "Number", "Page",
"www.rsc.org/dalton" );
$identAry['ChemicalCom']     = array( "Chemical Communications", "Number", "Page", "www.rsc.org/chemcomm" );
$identAry['ChemicalCom2']    = array( "View Article Online", "Table of Contents for th", "!*a", "!*b", "This article is part of
the", "www.rsc.org/chemcomm" );
$identAry['RoyalSocChem']    = array( "Title:", "rsc.li/catalysis", "Registered charity number" );           // "As featured in:
$identAry['BioMedCilia']     = array( "Cilia (", "DOI" );
$identAry['CellPress']   = array( "Article", "Graphical Abstract", "Authors", "Correspondence", "In Brief", "Highlights" /* ,
"Accession Numbers" */  );

$assistAry = array();        // to hold cover-type-specific information when the cover-page is over.
$assistAry['PhysChChPhys']['cover_end_terms'] = array( "www.rsc.org/pccp | Physical Chemistry Chemical Physics" );
$assistAry['PhysChChPhys']['cover_end_offset'] = -1;               // 0 to consider the page where cover_end_terms are found as
part of the cover.
$assistAry['DaltonTrans']['cover_end_terms'] = array( "www.rsc.org/dalton | Dalton Transactions" );
$assistAry['DaltonTrans']['cover_end_offset'] = -1;               // 0 to consider the page where cover_end_terms are found as
part of the cover.
$assistAry['DaltonTrans2']['cover_end_terms'] = array( "www.rsc.org/dalton | Dalton Transactions" );
$assistAry['DaltonTrans2']['cover_end_offset'] = -1;               // 0 to consider the page where cover_end_terms are found as
part of the cover.
$assistAry['ChemicalCom']['cover_end_terms'] = array( "www.rsc.org/chemcomm | ChemComm" );
$assistAry['ChemicalCom']['cover_end_offset'] = -1;               // 0 to consider the page where cover_end_terms are found as
part of the cover.
$assistAry['ChemicalCom2']['cover_end_terms'] = array( "www.rsc.org/chemcomm | ChemComm" );
$assistAry['ChemicalCom2']['cover_end_offset'] = -1;               // 0 to consider the page where cover_end_terms are found as
part of the cover.
$assistAry['WileyGeneral']['req_human_check'] = true;
$assistAry['BioMedCilia']['req_human_check'] = true;
```

From: Last, Sarah <Sarah.Last@lib4ri.ch>
Sent: Donnerstag, 28. September 2023 13:13
To: Konstantaki, Laura Amalia <Laura.Konstantaki@lib4ri.ch>; Hösli, Frank <Frank.Hoesli@lib4ri.ch>
Cc: Felder, Fabian <Fabian.Felder@lib4ri.ch>; Cantini, Federico <Federico.Cantini@lib4ri.ch>
Subject: RE: A task for Frank

Hi together

Regarding coverpages, I checked a few publications. Let me know if more publishers should be checked (e.g. IEEE; Cambridge, ASM, etc., so I would do an export of all publishers). Just checked now the most recent ones at the facets (cut at 100 publications, SPIE I added since I knew about their coverpages). I also found a very old e-mail, we have already checked once DORA for cover pages, I attach the mail.

Elsevier: no
Wiley: mostly not, but seems to be sometimes, atleast found one:
https://onlinelibrary.wiley.com/doi/epdf/10.1002/adfm.201101830
Springer Nature: no
American Chemical Society: no

Taylor & Francis: no
IOP Publishing: yes, example: https://iopscience.iop.org/article/10.1088/0957-4484/23/25/255705
American Institute of Physics: yes, example: https://pubs.aip.org/avs/sss/article/19/1/62/366303/The-Si3N4-TiN-Interface-4-Si3N4-TiN-001-Grown-with
Copernicus: no
Royal Society of Chemistry: mostly not, but seems to be sometimes, found one: https://pubs.rsc.org/en/content/articlelanding/2011/CC/c1cc12490k
Springer: no
SPIE: yes, e.g. https://doi.org/10.1117/12.918246

LG
Sarah

From: Konstantaki, Laura Amalia <Laura.Konstantaki@lib4ri.ch>
Sent: Donnerstag, 28. September 2023 11:28
To: Hösli, Frank <Frank.Hoesli@lib4ri.ch>; Last, Sarah <Sarah.Last@lib4ri.ch>
Cc: Felder, Fabian <Fabian.Felder@lib4ri.ch>; Cantini, Federico <Federico.Cantini@lib4ri.ch>
Subject: RE: A task for Frank

Hi Frank,

Thanks alot for checking! Please see below some comments:

From: Hösli, Frank <Frank.Hoesli@lib4ri.ch>
Sent: Mittwoch, 27. September 2023 12:48
To: Last, Sarah <Sarah.Last@lib4ri.ch>
Cc: Konstantaki, Laura Amalia <Laura.Konstantaki@lib4ri.ch>; Felder, Fabian <Fabian.Felder@lib4ri.ch>; Cantini, Federico <Federico.Cantini@lib4ri.ch>
Subject: FW: A task for Frank

Hi Together

http://lib-dora-dev1.emp-eaw.ch:3000/issues/318 :
>> Jan 2023: Also check the problem from Eawag reports ingested -> some PDFA conversion did not work properly.

I have left a remark in the ticked what I have found out so far about this matter.

Kind Regards
Frank

From: Hösli, Frank
Sent: Dienstag, 26. September 2023 16:02
To: Last, Sarah <Sarah.Last@lib4ri.ch>
Cc: Konstantaki, Laura Amalia <Laura.Konstantaki@lib4ri.ch>; Felder, Fabian <Fabian.Felder@lib4ri.ch>; Cantini, Federico <Federico.Cantini@lib4ri.ch>
Subject: RE: A task for Frank

Hi Together

For the majority of these case we can use routines of Federico's PDF-Downloader https://github.com/Lib4RI/pdf_downloader/ (as used within the Scopus Alert).
There should not be nuch problem integreting this into a batch job where we deal with the publications/MODS/PDFs.

Great!

By default the PDF-Downloader needs a DOI, and among the totally 8596 publications, 5751 publications have a DOI already:
http://lib-dora-prod1.emp-eaw.ch:8080/solr/collection1/select?wt=csv&indent=true&csv.separator=|&sort=PID+asc&rows=987654321&q=PID:*%5c%3a*+AND+mods_identifier_doi_mt:10*+AND+mods_originInfo_encoding_w3cdtf_keyDate_yes_dateIssued_dt:[2006-01-01T00%3A00%3A00Z%20TO%202012-01-01T00%3A00%3A00Z]+NOT+(RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/psi%5c%3Apublications%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/psi%5c%3Aexternal%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Apublications%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Awsl%5c-int%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Awsl%5c-ext%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/eawag%5c%3Aext%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Aforum%22)&fl=PID%2c+mods_identifier_doi_mt%2c+mods_identifier_doi_mt%2c+mods_identifier_scopus_mt%2c+mods_identifier_ut_mt%2c+mods_note_notesLib4RI_mt%2c+mods_note_additional?information_mt

Great!

And for 456 publications there is no DOI though, but a Scopus or WoS ID:
http://lib-dora-prod1.emp-eaw.ch:8080/solr/collection1/select?wt=csv&indent=true&csv.separator=|&sort=PID+asc&rows=987654321&q=PID:*%5c%3a*+NOT+mods_identifier_doi_mt:10*+AND+(mods_identifier_scopus_mt:*+OR+mods_identifier_ut_mt)+AND+mods_originInfo_encoding_w3cdtf_keyDate_yes_dateIssued_dt:[2006-01-01T00%3A00%3A00Z%20TO%202012-01-01T00%3A00%3A00Z]+NOT+(RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/psi%5c%3Apublications%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/psi%5c%3Aexternal%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Apublications%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Awsl%5c-int%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Awsl%5c-ext%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/eawag%5c%3Aext%22+OR+RELS_EXT_isMemberOfCollection_uri_ms:%22info%5c%3Afedora%5c/wsl%5c%3Aforum%22)&fl=PID%2c+mods_identifier_doi_mt%2c+mods_identifier_doi_mt%2c+mods_identifier_scopus_mt%2c+mods_identifier_ut_mt%2c+mods_note_notesLib4RI_mt%2c+mods_note_additional?information_mt

Via the correponding Elsevier/Clarivate APIs we could try to get a DOI as well (which we also could write back into the MODS if you like, aside from removing '#check_pdf' if there is).

I am not sure if we should do this, sometimes DOI's saved in WoS e.g. are wrong and we have removed them in purpose. I would not do this extra step.

And the ~2500 publications without any ID - we can identify/log them at least...
No need to log them.

But concerning an automatic download/replacement...:
- How could we know/estmate the new PDF is better than the old one?
Lothar remembers that we did some bad editing on the PDFS during th