unpublished: Hercher, J.; Sack, H. & Waitelonis, J.: Building a Research Context with Linked Bibliographic Data, contribution to OFKN Open Bibliographic Data Challenge, Feb. 18, 2011

Building a Research Context with Linked Bibliographic Data

Johannes Hercher, Harald Sack, and Joerg Waitelonis

Hasso-Plattner-Institut für Softwaresystemtechnik GmbH,
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany
{johannes.hercher,harald.sack,joerg.waitelonis}@hpi.uni-potsdam.de
http://www.hpi.uni-potsdam.de

This proposed service is designed to advance learning and research by taking bibliographic datasets as backbone to gather, interlink, and suggest related information, i.e. to build an research context that enables a quick overview on a topic by the use of Linked Bibliographic Data and explorative search techniques. The general idea is to recommend appropriate literature and media based on a bibliographic entry of the *IUCr-Dataset*[1]. These IUCr Documents contain author, title and abstract information (the bibset) that is sufficient to find related wikipedia articles. In order to accomplish this in a most convenient way we harness free semantic webservices, as e.g., *Zemanta*[2] and *Open Calais*[3]. For Instance the metadata of the IUCr Document *N-(2-Ethylphenyl)phthalimide*[1] lead to the scientific Wikipedia articles *phthalimide*, *dihedral angle*, and *benzene*. Table 1 shows the original metadata from an IUCr document and gathered metadata from external webservices. While Zemanta delivers useful links to Wikipedia pages, Open Calais provides keywords that can be harnessed for ranking issues, cf. fig. 1 and tab. 1.

Original IUCr document	**Title**: N-(2-Ethylphenyl)phthalimide **Authors**: Y. M. Fan, N. Zakaria, A. Ariffin and S. W. Ng **Abstract**: In the title compound, C16H13NO2, the phthalimide and benzene ring systems form a dihedral angle of 77.2 (1).
Zemanta's Wikilinks	http://en.wikipedia.org/wiki/Dihedral_angle http://en.wikipedia.org/wiki/Phthalimide http://en.wikipedia.org/wiki/Benzene
Open Calais' Keywords	Organic chemistry, Mathematics, Chemistry, Dihedral, Benzene, Phthalimide, Simple aromatic rings, Imides, Aromatic compounds

Table 1. Retrieved metadata from Open Calais and Zemanta for an IUCr document [1].

[1] The bibliographic data from Acta Cryst E, a publication by the International Union of Crystallography (IUCr), cf. http://ckan.net/package/iucr-acta-cryst- eInternational Union of Crystallography (IUCr)
[2] http://www.zemanta.com/api and http://www.zemanta.com/demo
[3] http://viewer.opencalais.com/

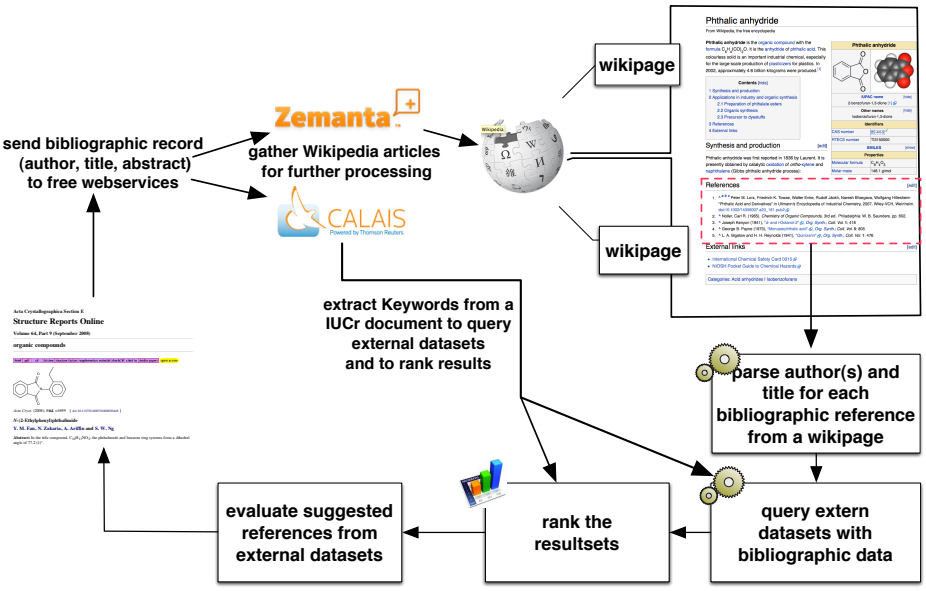

Fig. 1. A proposed workflow to augument IUCr documents with recommendations from external databases.

We proceed with Wikipedia articles that can be scanned for bibliographic information (aka the reference section). Pages without references are withdrawn, because we assume that they are not scientific enough nor relevant. Widely consistent HTML-patterns make it possible to simply extract bibliographic references from the wikipedia (the wikisets), in an automated way. Wikisets, i.e. author and title information, are used to query open datasets, as e.g., British National Bibliography (BNL)[4], the Cambridge University Library (CUL)[5], the IUCr itselfe, and many others. The gathered result-sets are ranked based on same matching author names and similarities in title-text. We assume that this tool will be helpful, either for shortening the process of interdisciplinary information research as well as for young researchers that may not be aware of free information sources available in their field.

To prove the feasibility of our idea, we examined six of 70 bibliographic references that where harvested by manual execution of the proposed workflow above, and shall relate to the example document [1]. Most of the harvested references come from an article on *benzene*, thus each of the articles on *dihedral angle* and *phthalimide* provided only three citations. Taking these references into account we query the catalogue of the British National Library (BNL)[6] and the

[4] http://ckan.net/package/jiscopenbib-bl_bnb-1
[5] http://ckan.net/package/jiscopenbib-cul-1
[6] http://catalogue.bl.uk/ (integrated catalogue)

Cambridge University Library[7] manually to compare them with results from IUCr[8] (tab. 2).

Harvested Citation	**BNL**	**IUCr**	**CUL**
Olshevsky, George, Dihedral angle at Glossary for Hyperspace. (online)	0/0/1	0/0/3	0/0/0
Weisstein, Eric W., Dihedral angle from MathWorld. (online)	0/0/1	0/0/3	1/0/0
David R. Lide, ed.. Physical Constants of Organic Compounds, in CRC Handbook of Chemistry and Physics, Internet Version 2005, <http://www.hbcpnetbase.com>. CRC Press.	28/0/5	1/0/59	13/0/5
O. T. Benfey, August Kekul and the Birth of the Structural Theory of Organic Chemistry in 1858, Journal of Chemical Education, 35 (1958), 21–23	1/0/5	0/0/21	1/0/320
Moran, Damian; Simmonett, Andrew C.; Leach, Franklin E.; Allen, Wesley D.; Schleyer, Paul v. R.; Schaefer, Henry F. (2006). Popular Theoretical Methods Predict Benzene and Arenes To Be Nonplanar. Journal of the American Chemical Society 128 (29): 9342. doi:10.1021/ja0630285. PMID 16848464.	0/0/2	1/0/0	0/0/0
Stranks, D. R.; M. L. Heffernan, K. C. Lee Dow, P. T. McTigue, G. R. A. Withers (1970). Chemistry: A structural view. Carlton, Victoria: Melbourne University Press. pp. 347. ISBN 0 522 83988 6.	4/15/15	0/0/0	5/7/7
+ 64 harvested citations			

Table 2. Selected references taken from Wikipedia articles. The columns show query results from British National Library (BNL), Crystallography Journals Online (IUCr) and the Cambridge University Library (CUL) in the manner of records from same author / records with same title / records with similar title.

We found 53 bibliographic references of the same author (33 in BNL and 20 in CUL), whereas we considered only the first author of a publication. Additional we found 22 bibliographic references, which had an exact overlap in the title of the bibliographic entry (15 in BNL and 7 in CUL). We also gathered 362 bibliographic entries (30 BNL and 332 in CUL) that have a similar title to the gathered Wikipedia citations from Table 2. In contrast we found no additional document of *Y. M. Fan* in the IUCr Dataset, but 10 documents with a similar title. These results show that it would be valuable to connect bibliographic

[7] http://ul-newton.lib.cam.ac.uk/vwebv/searchBasic?sk=en_US
[8] http://journals.iucr.org/

datasets in general, but we encountered some difficulties as well. Short Wikipedia articles usally do not use citation templates[9]. Its also possible to retrieve irrelevant sources due to ambiguous terms and different authors with the same name. In future we will consider extraction and normalization of bibliographic information from the wikipedia in order to connect these references to normative authority files of public libraries. The recommendation of interdisciplinary resources may be also improved by taking identifiers, as e.g., the CAS-Number or PubChem-Nr. from the infoboxes of Wikipedia into account.

References

1. Fan, Y.M., Zakaria, N., Ariffin, A., Ng, S.W.: *N*-(2-Ethylphenyl)phthalimide. Acta Crystallographica Section E **64**(9) (Sep 2008) o1699

[9] http://en.wikipedia.org/wiki/Wikipedia:Citation_templates