



AUTOMATING BIG DATA CLEANING: AN EXAMPLE USING LOCAL BIBLIOMETRIC DATA

Jana Carson, Manager, Evaluation & Accountability, Institutional Analysis & Planning
Shannon Gordon, Bibliometrics & Research Impact Librarian, Library

December 6, 2016

watitis.uwaterloo.ca
[@watitiscnf](https://twitter.com/watitiscnf)



#watitis2016



AGENDA

- Background
- Project
- Next steps
- Discussion



#watitis2016

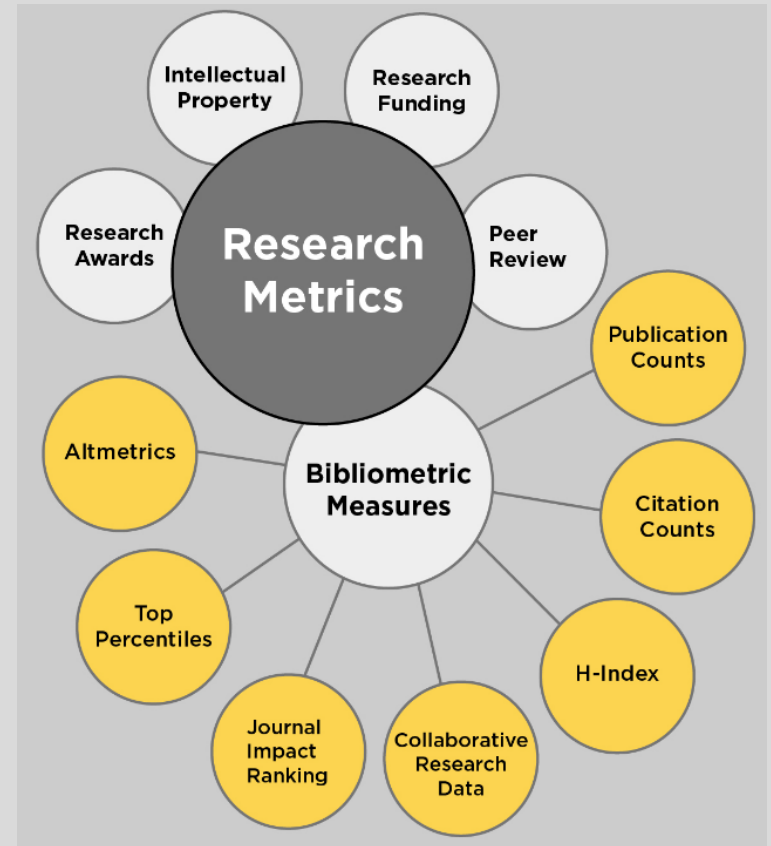


UNIVERSITY OF
WATERLOO

THE PROJECT

- Research area
 - » Strategic highlights
 - » Research funding
 - » Research productivity

(University of Waterloo Working Group on Bibliometrics, 2016)



THE DATA



#watitis2016



UNIVERSITY OF
WATERLOO

RESEARCH PRODUCTIVITY

Total publications
(annual, 5-year,
overall)

Total citations
(annual, 5-year,
overall)

Average #
citations per
publication

Median number of
citations

Average #
publications per
member

Total publications
with citations

Total publications
with zero citations

Total publications
with 100 or more
citations

H-index

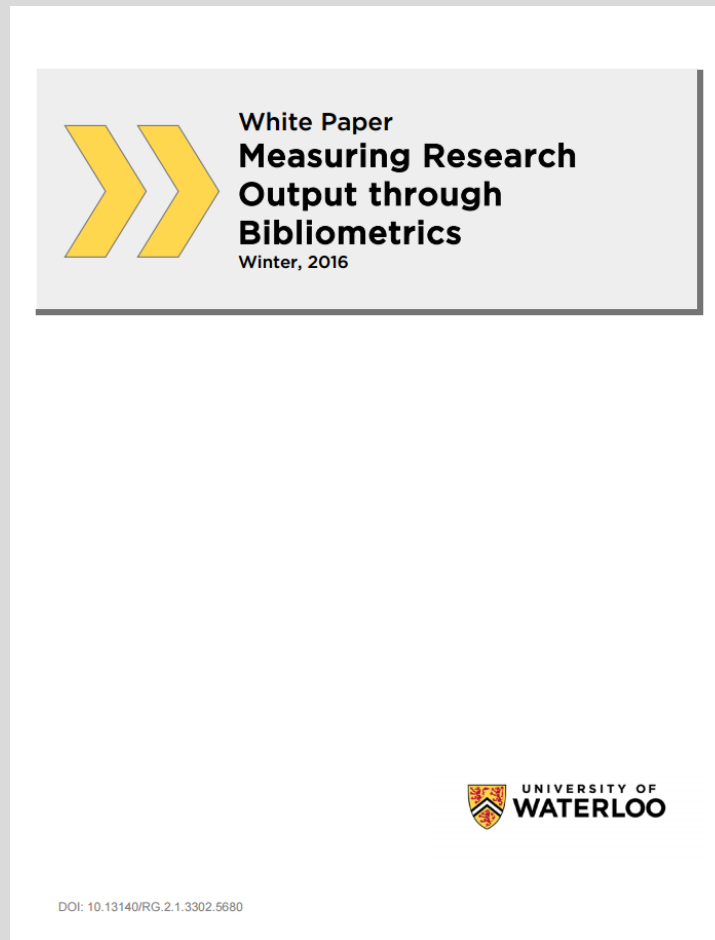


#watitis2016



UNIVERSITY OF
WATERLOO

WHAT GUIDES OUR WORK?



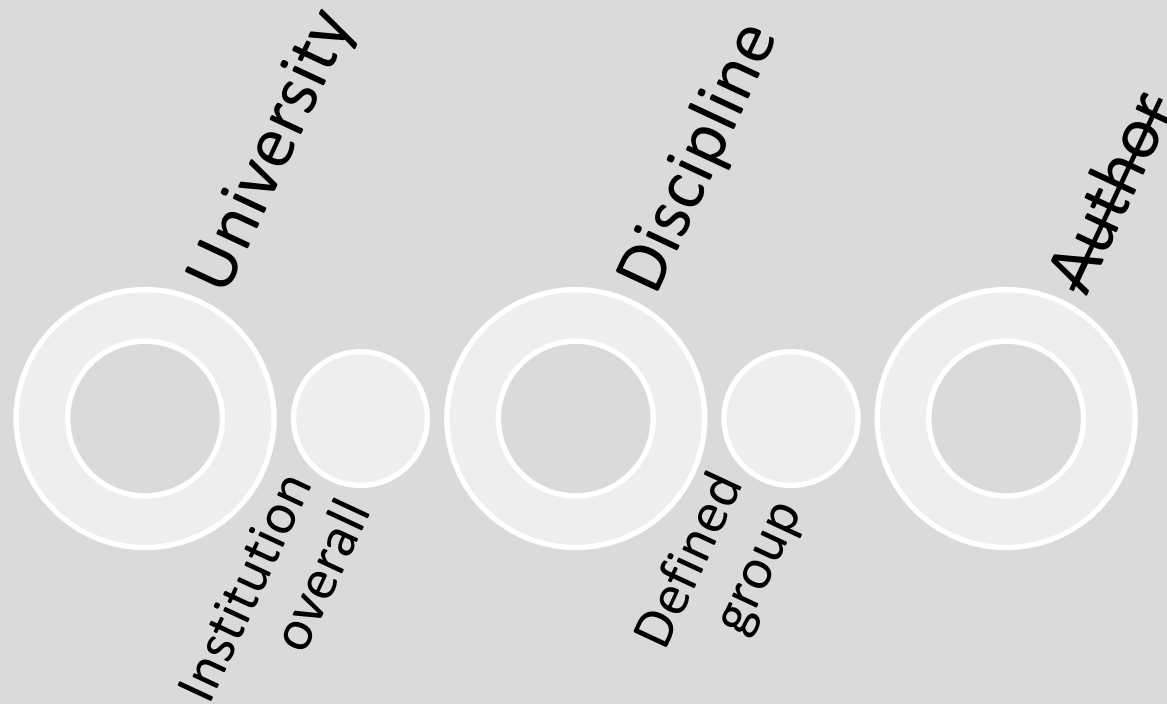
#watitis2016

The logo for WatITis, featuring the word "WatITis" in a bold, black, sans-serif font. The "IT" is in a larger, bolder font. Above the "IT" is a red arch that spans the width of the "IT" and extends slightly beyond it on both sides.



UNIVERSITY OF
WATERLOO

INTERNAL ANALYSIS



#watitis2016



UNIVERSITY OF
WATERLOO

DATA SOURCE

The image displays two overlapping screenshots of Thomson Reuters' research databases. The background screenshot is the 'WEB OF SCIENCE™' interface, featuring a search bar with the example text 'Example: oil spill* mediterranean', a 'Basic Search' dropdown, and a 'TIMESPAN' section with radio buttons for 'All years' and 'From 1900 to 2015'. The foreground screenshot is the 'InCites™' interface, titled 'Calibrate Your Strategic Research Vision'. It shows a 'New Tile' dashboard with a 'Results: 30,678,221' count and a 'Geographic' filter. A world map is displayed, with countries colored in shades of orange and red, indicating research density. The map shows high concentrations in North America, Europe, and East Asia. The InCites interface also includes a 'Dataset' section with 'InCites Dataset' selected, and a 'Filters' section with options for 'By Attributes' (Person Name or ID, Affiliated Organization, Location) and 'By Research Network'.



#watitis2016

WatITis



UNIVERSITY OF
WATERLOO

OUR RAW DATA...EEK!

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Accession Number	DOI	Pubmed ID	Article Title	Link	Authors	Source	Volume	Issue	Pages	Publication Date	Times Cited	Journal Expected	Category Expected	Journal Normalized	Category Normalized	Percentile in Subject	Journal Impact
2	WOS:0003373502	10.1038/nature.2014.13032	25111111	Contextuality sup	https://doi.org/10.1038/nature.2014.13032	Howard, Mark; Wa	NATURE	510	7505	351-355	2014	27	38.58	3.29	0.7	8.21	1.37	41.46
3	WOS:0003440312	10.1103/PhysRevLett.112.110301	25111111	Interferometric P	https://doi.org/10.1103/PhysRevLett.112.110301	Serbyn, M.; Knap	PHYSICAL REVIEW LETTERS	113	14		2014	27	9.85	3.29	2.74	8.21	1.37	7.51
4	WOS:0003323298	10.1103/PhysRevA.89.042101	25111111	Improved entropi	https://doi.org/10.1103/PhysRevA.89.042101	Coles, Patrick J.; P	PHYSICAL REVIEW A	89	2		2014	18	3.35	2.9	5.37	6.21	0.96	2.81
5	WOS:0003305748	10.1088/1367-2630/16/10/103001	25111111	The resource the	https://doi.org/10.1088/1367-2630/16/10/103001	Veitch, Victor; Mo	NEW JOURNAL OF PHYSICS	16			2014	18	4.19	3.29	4.3	5.47	3.06	3.56
6	WOS:0003331882	10.1103/PhysRevA.89.042101	25111111	Casimir forces on	https://doi.org/10.1103/PhysRevA.89.042101	Alhambra, Alvaro	PHYSICAL REVIEW A	89	3		2014	17	3.35	2.9	5.07	5.87	1.08	2.81
7	WOS:0003382844	10.1126/science.1251111	25111111	Tunable fractiona	https://doi.org/10.1126/science.1251111	Maier, Patrick; W	SCIENCE	345	6192	61-64	2014	17	30.97	3.29	0.55	5.17	3.52	33.61
8	WOS:0003351567	10.1007/s00220-014-1911-1	25111111	Everything You Al	https://doi.org/10.1007/s00220-014-1911-1	Chitambar, Eric; L	COMMUNICATIONS IN MATHEMATICAL PHYSICS	328	1	303-326	2014	15	3.15	2	4.76	7.51	1.03	2.09
9	WOS:0003338000	10.1038/NPHOTON.2014.010101	25111111	Experimental thre	https://doi.org/10.1038/NPHOTON.2014.010101	Erven, C.; Meyer-S	NATURE PHOTONICS	8	4	292-296	2014	12	31.57	2.89	0.38	4.16	2.7	32.39
10	WOS:0003319396	10.1103/PhysRevLett.112.110301	25111111	Using Concatenat	https://doi.org/10.1103/PhysRevLett.112.110301	Jochym-O'Connor	PHYSICAL REVIEW LETTERS	112	1		2014	12	9.85	3.29	1.22	3.65	6.28	7.51
11	WOS:0003431452	10.1038/NPHOTON.2014.010101	25111111	Direct generation	https://doi.org/10.1038/NPHOTON.2014.010101	Hamel, Deny R.; S	NATURE PHOTONICS	8	10	801-807	2014	11	31.57	2.89	0.35	3.81	3.39	32.39
12	WOS:0003359288	10.1103/PhysRevLett.112.110301	25111111	Hamiltonian Lear	https://doi.org/10.1103/PhysRevLett.112.110301	Wiebe, Nathan; G	PHYSICAL REVIEW LETTERS	112	19		2014	11	9.85	3.29	1.12	3.35	7.24	7.51
13	WOS:0003319525	10.1103/PhysRevLett.112.110301	25111111	Laser Damage Hel	https://doi.org/10.1103/PhysRevLett.112.110301	Bugge, Audun Nys	PHYSICAL REVIEW LETTERS	112	7		2014	10	9.85	3.29	1.02	3.04	8.38	7.51
14	WOS:0003345931	10.1103/PhysRevA.89.042101	25111111	Quantum Hamilto	https://doi.org/10.1103/PhysRevA.89.042101	Wiebe, Nathan; G	PHYSICAL REVIEW A	89	4		2014	10	3.35	2.9	2.99	3.45	4.15	2.81
15	WOS:0003372777	10.1103/PhysRevLett.112.110301	25111111	Experimental Rea	https://doi.org/10.1103/PhysRevLett.112.110301	Li, Zhaokai; Zhou	PHYSICAL REVIEW LETTERS	112	22		2014	9	9.85	3.29	0.91	2.74	9.7	7.51
16	WOS:0003463228	10.1021/nl5c00000a001	25111111	Polarization Enta	https://doi.org/10.1021/nl5c00000a001	Huber, Tobias; Pre	NANO LETTERS	14	12	7107-71	2014	9	14.57	4.62	0.62	1.95	9.12	13.59
17	WOS:0003444068	10.1103/PhysRevB.90.100101	25111111	Quantum quench	https://doi.org/10.1103/PhysRevB.90.100101	Serbyn, Maksym; F	PHYSICAL REVIEW B	90	17		2014	9	4.35	4.5	2.07	2	12.98	3.74
18	WOS:0003440961	10.1088/1367-2630/16/10/103001	25111111	Randomized bench	https://doi.org/10.1088/1367-2630/16/10/103001	Wallman, Joel J.; F	NEW JOURNAL OF PHYSICS	16			2014	9	4.19	3.29	2.15	2.74	9.7	3.56
19	WOS:0003319468	10.1103/PhysRevLett.112.110301	25111111	Topological Phase	https://doi.org/10.1103/PhysRevLett.112.110301	Papic, Z.; Abanin	PHYSICAL REVIEW LETTERS	112	4		2014	9	9.85	3.29	0.91	2.74	9.7	7.51
20	WOS:0003412607	10.1103/PhysRevA.90.042101	25111111	Nonclassical high	https://doi.org/10.1103/PhysRevA.90.042101	Rundquist, Arman	PHYSICAL REVIEW A	90	2		2014	8	3.35	2.9	2.39	2.76	6.89	2.81
21	WOS:0003321690	10.1103/PhysRevLett.112.110301	25111111	Quantum Nonder	https://doi.org/10.1103/PhysRevLett.112.110301	Sathiamoorthy, S	PHYSICAL REVIEW LETTERS	112	9		2014	8	9.85	3.29	0.81	2.43	11.27	7.51
22	WOS:0003323306	10.1103/PhysRevA.89.042101	25111111	Quantum signalin	https://doi.org/10.1103/PhysRevA.89.042101	Jonsson, Robert H.	PHYSICAL REVIEW A	89	2		2014	8	3.35	2.9	2.39	2.76	6.89	2.81
23	WOS:0003321680	10.1103/PhysRevD.90.042101	25111111	Echo of the quant	https://doi.org/10.1103/PhysRevD.90.042101	Garay, Luis J.; Mar	PHYSICAL REVIEW D	89	4		2014	7	6.22	5	1.12	1.4	22.11	4.64
24	WOS:0003446073	10.1088/0264-9381/31/10/103001	25111111	Entanglement in c	https://doi.org/10.1088/0264-9381/31/10/103001	Martin-Martinez, E	CLASSICAL AND QUANTUM GRAVITY	31	21		2014	7	3.68	4.26	1.9	1.64	13.39	3.17
25	WOS:0003388438	10.1088/1367-2630/16/10/103001	25111111	Experimental real	https://doi.org/10.1088/1367-2630/16/10/103001	Lu, Dawei; Broduri	NEW JOURNAL OF PHYSICS	16			2014	7	4.19	3.29	1.67	2.13	13.39	3.56
26	WOS:0003345133	10.1088/1367-2630/16/10/103001	25111111	Memory-assisted	https://doi.org/10.1088/1367-2630/16/10/103001	Panayi, Christiana	NEW JOURNAL OF PHYSICS	16			2014	7	4.19	3.29	1.67	2.13	13.39	3.56
27	WOS:0003394461	10.1103/PhysRevB.90.042101	25111111	Single-mode app	https://doi.org/10.1103/PhysRevB.90.042101	Repellin, Cecile; N	PHYSICAL REVIEW B	90	4		2014	7	4.35	4.5	1.61	1.56	18.04	3.74
28	WOS:0003312885	10.1103/PhysRevD.90.042101	25111111	Spacetime effects	https://doi.org/10.1103/PhysRevD.90.042101	Bruschi, David Edw	PHYSICAL REVIEW D	90	4		2014	7	6.22	5	1.12	1.4	22.11	4.64
29	WOS:0003386343	10.1103/PhysRevLett.112.110301	25111111	Ultrafast and Faul	https://doi.org/10.1103/PhysRevLett.112.110301	Muralidharan, Sri	PHYSICAL REVIEW LETTERS	112	25		2014	7	9.85	3.29	0.71	2.13	13.39	7.51
30	WOS:0003343004	10.1038/ncomms25111	25111111	An experimental	https://doi.org/10.1038/ncomms25111	Erven, C.; Ng, N.; C	NATURE COMMUNICATIONS	5			2014	6	10.31	2.47	0.58	2.43	12.03	11.47
31	WOS:0003476137	10.1038/ncomms25111	25111111	Equivalence of w	https://doi.org/10.1038/ncomms25111	Coles, Patrick J.; K	NATURE COMMUNICATIONS	5			2014	6	10.31	3.29	0.58	1.82	15.89	11.47
32	WOS:0003369153	10.1103/PhysRevA.89.042101	25111111	Quantum fingerpi	https://doi.org/10.1103/PhysRevA.89.042101	Arrazola, Juan Mig	PHYSICAL REVIEW A	89	6		2014	6	3.35	2.9	1.79	2.07	12.03	2.81
33	WOS:0003331072	10.1103/PhysRevD.89.042101	25111111	Scalar Green func	https://doi.org/10.1103/PhysRevD.89.042101	Yang, Huan; Zhang	PHYSICAL REVIEW D	89	6		2014	6	6.22	5	0.96	1.2	26.41	4.64
34	WOS:0003412370	10.1103/PhysRevB.90.042101	25111111	Solvable models f	https://doi.org/10.1103/PhysRevB.90.042101	Papic, Z.	PHYSICAL REVIEW B	90	7		2014	6	4.35	4.5	1.38	1.33	21.63	3.74
35	WOS:0003322216	10.1103/PhysRevA.89.042101	25111111	Tractable simulat	https://doi.org/10.1103/PhysRevA.89.042101	Puzzuoli, Daniel; C	PHYSICAL REVIEW A	89	2		2014	6	3.35	2.9	1.79	2.07	12.03	2.81



#watitis2016



UNIVERSITY OF
WATERLOO

DATA STAGES



#watitis2016



UNIVERSITY OF
WATERLOO

INPUT DATA WORKFLOW



Author A
Publication 1
Publication 2
Publication 3



OUTPUT DATA WORKFLOW



Year A
Publication 1
Publication 2
Publication 3



#watitis2016

THE PROBLEM



Isolating defined data points

Populating author-level publication
worksheets

Updating needed data points

Calculating various data points

Cleaning up all data & generating final
datasets



#watitis2016



UNIVERSITY OF
WATERLOO

FINDING A SOLUTION



#watitis2016



UNIVERSITY OF
WATERLOO

What processes did we automate?



#watitis2016



UNIVERSITY OF
WATERLOO

DATA PROBLEM #1

Problem	Manual process	Automated solution	Time savings
Isolating relevant author names in a publication dataset	Rigorous manual task 18-20 hours per research area	Automated Python script automatically isolates relevant author name variants	Significant! 108-120 manual hours



#watitis2016



UNIVERSITY OF
WATERLOO

ISOLATING DEFINED DATA POINTS

	J	K	L	M	N	O	P	Q	R	S
1	Publication	Subject Area	Document Type	All Authors	Journal	Document Title	Unique Identifier	Volume	Page	uW Author(s)
2	2004	ENGINEERING	ARTICLE	Shao, P; HUANG, Y	AIChE JOURNAL	Gas-liquid distillation	2.204E+11	50	557-565	Relevant uW authors isolated for each publication
3	2004	ENVIRONMENTAL	ARTICLE	Dorner, SM; Huck, J	JOURNAL OF	Assessing levels of	2.242E+11	67	1813-1823	
4	2005	ENGINEERING	ARTICLE	Shao, P; HUANG, Y	JOURNAL OF	Composite materials	2.295E+11	254	#####	
5	2005	ENGINEERING	ARTICLE	Pintar, KDM; ANDERSON, W	JOURNAL OF	Assessment of	2.304E+11	97	116-129	
6	2006	ENVIRONMENTAL	ARTICLE	Dorner, Sarah M.;	ENVIRONMENTAL	Hydrologic modeling	2.394E+11	40	4746-4753	
7	2006	WATER RESOURCES	ARTICLE	Ndiongue, Souleymane	WATER QUALITY	Using pilot-scale	2.421E+11	41	296-306	
8	2006	ENGINEERING	PROCEEDING	Espino-Cortes, F.;	2006 Annual	Water processing	2.425E+11		684-687	
9	2007	WATER RESOURCES	ARTICLE	Dorner, Sarah M.;	JOURNAL OF	Pathogen analysis	2.468E+11	5	241-257	
10	2007	FOOD SCIENCE	ARTICLE	Alani, Faiez; Moores, G	FOOD BIOTECHNOLOGY	Optimization of	2.478E+11	21	169-180	
11	2007	WATER RESOURCES	ARTICLE	Anderson, William	JOURNAL OF	Estimation of	2.511E+11	5	553-572	
12	2007	ENVIRONMENTAL	ARTICLE	Zhang, Jianping (UNIV OF	WATER				451-460	
13	2008	ENGINEERING	ARTICLE	Zhang, Jianping (UNIV OF	JOURNAL OF				92	
14	2008	ENGINEERING	ARTICLE	Anderson, William	JOURNAL OF				585-597	
15	2009	WATER RESOURCES	LETTER	Anderson, William	JOURNAL OF	THINKING OF	2.645E+11	101	16	
16	2009	WATER RESOURCES	ARTICLE	Cheyne, Bo M. (UNIV OF	JOURNAL OF	An evaluation of	2.696E+11	7	392-403	
17	2009	WATER RESOURCES	REVIEW	Hamouda, M. A. (UNIV OF	WATER SCIENCE	Decision support	2.709E+11	60	1757-1770	
18	2009	ENGINEERING	ARTICLE	Anderson, William	JOURNAL OF	Influence of	2.715E+11	58	450-462	
19	2010	WATER RESOURCES	ARTICLE	Cheyne, Bo M. (UNIV OF	JOURNAL OF	The detection of	2.818E+11	8	487-499	
20	2010	ENGINEERING	ARTICLE	Hamouda, Mohamad	JOURNAL OF	A framework for	2.86E+11	102	42-56	
21	2011	WATER RESOURCES	ARTICLE	Zhang, Jianping (HONG KONG	JOURNAL OF	A comparison of	2.913E+11	60	197-209	
22	2012	BIOTECHNOLOGY	ARTICLE	Alani, Faiez (MCMaster	WORLD JOURNAL	Biosynthesis of	3.003E+11	28	1081-1086	
23	2013	ENVIRONMENTAL	ARTICLE	Chen, Fei (UNIVERSITY OF	CHEMOSPHERE	An approach to	3.126E+11	90	758-765	
24	2013	ENVIRONMENTAL	ARTICLE	Chen, Fei (UNIVERSITY OF	ENVIRONMENTAL	Application of	3.207E+11	179	224-231	
25	2012	WATER RESOURCES	ARTICLE	Hamouda, M. A. (UNIV OF	WATER SCIENCE	Employing m	3.277E+11	12	637-647	
26	2014	WATER RESOURCES	REVIEW	Rahman, Mohamad	WATER RESEARCH	Behaviour and	3.309E+11	50	318-340	
27	2013	ENGINEERING	ARTICLE	Chen, Fei (UNIVERSITY OF	WATER SCIENCE	An approach to	3.343E+11	13	835-845	
28	2014	PUBLIC ENVIRONMENTAL	ARTICLE	Hammond, David	EUROPEAN	The percepti	3.395E+11	24	527-542	



#watitis2016



UNIVERSITY OF
WATERLOO

DATA PROBLEM #2

Problem	Manual process	Automated solution	Time savings
Populating author-level publication worksheets for each author affiliated with a research area	Manual process involving much copying/pasting 4-6 hours per area	Python script automatically generates relevant author-level tab in target file	24-36 manual hours



#watitis2016



UNIVERSITY OF
WATERLOO

POPULATING WORKSHEETS WITH DEFINED DATA POINTS

	J	K	L	M	N	O	P	Q	R	S
1	Publication	Subject Area	Document Type	All Authors	Journal	Document Title	Unique Identifier	Volume	Page	uW Author(s)
2	2004	ENGINEERING	ARTICLE	Shao, P; HUANG, A	AICHE JOURNAL	Gas-liquid distillation	2.204E+11	50	557-565	Worksheet lists all publications authored by Author A
3	2004	ENVIRONMENTAL	ARTICLE	Dorner, SM; Huck	JOURNAL OF	Assessing levels of	2.242E+11	67	1813-1823	
4	2005	ENGINEERING	ARTICLE	Shao, P; HUANG, A	JOURNAL OF	Composite materials	2.295E+11	254	#####	
5	2005	ENGINEERING	ARTICLE	Pintar, KDM; AND	JOURNAL OF	Assessment of	2.304E+11	97	116-129	
6	2006	ENVIRONMENTAL	ARTICLE	Dorner, Sarah M.;	ENVIRONMENTAL	Hydrologic modeling	2.394E+11	40	4746-4753	
7	2006	WATER RESOURCES	ARTICLE	Ndiongue, Souley	WATER QUALITY	Using pilot-scale	2.421E+11	41	296-306	
8	2006	ENGINEERING	PROCEEDINGS	Espino-Cortes, F.;	2006 Annual	Water processing	2.425E+11		684-687	
9	2007	WATER RESOURCES	ARTICLE	Dorner, Sarah M.;	JOURNAL OF	Pathogen analysis	2.468E+11	5	241-257	
10	2007	FOOD SCIENCE	ARTICLE	Alani, Faiez; Moo	FOOD BIOTECHNOLOGY	Optimization of	2.478E+11	21	169-180	
11	2007	WATER RESOURCES	ARTICLE	Anderson, William	JOURNAL OF	Estimation of	2.511E+11	5	553-572	
12	2007	ENVIRONMENTAL	ARTICLE	Zhang, Jianping (U	OZONE-SCIENCE	A computational	2.524E+11	29	451-460	
13	2008	ENGINEERING	ARTICLE	Zhang, Jianping (U	JOURNAL OF	Application of	2.538E+11	57	79-92	
14	2008	ENGINEERING	ARTICLE	Anderson, William	JOURNAL OF	Endotoxin release	2.612E+11	57	585-597	
15	2009	WATER RESOURCES	LETTER	Anderson, William	JOURNAL OF	THINKING OF	2.645E+11	101	16	
16	2009	WATER RESOURCES	ARTICLE	Cheyne, Bo M. (U	JOURNAL OF	An evaluation of	2.696E+11	7	392-403	
17	2009	WATER RESOURCES	REVIEW	Hamouda, M. A. (	WATER SCIENCE	Decision support	2.709E+11	60	1757-1770	
18	2009	ENGINEERING	ARTICLE	Anderson, William	JOURNAL OF	Influence of	2.715E+11	58	45-462	
19	2010	WATER RESOURCES	ARTICLE							
20	2010	ENVIRONMENTAL	ARTICLE							
21	2011	WATER RESOURCES	ARTICLE	Zhang, Jianping (H	JOURNAL OF	A comparison of	2.913E+11	60	19-209	
22	2011	TECHNOLOGY	ARTICLE	Alani, Faiez (MCM	WORLD JOURNAL	Biosynthesis of	3.003E+11	28	1081-1086	
23	2011	ENVIRONMENTAL	ARTICLE	Chen, Fei (UNIV V	CHEMOSPHERE	An approach to	3.126E+11	90	758-765	
24	2011	ENVIRONMENTAL	ARTICLE	Chen, Fei (UNIV V	ENVIRONMENTAL	Application of	3.207E+11	179	224-231	
25	2011	WATER RESOURCES	ARTICLE	Hamouda, M. A. (	WATER SCIENCE	Employing m	3.277E+11	12	637-647	
26	2011	WATER RESOURCES	REVIEW	Rahman, Moham	WATER RESEARCH	Behaviour and	3.309E+11	50	318-340	
27	2013	ENGINEERING	ARTICLE	Chen, Fei (UNIV V	WATER SCIENCE	An approach to	3.343E+11	13	835-845	



#watitis2016



UNIVERSITY OF
WATERLOO

DATA PROBLEM #3

Problem	Automated solution
3A: Update 2004-2013 publication data with new metrics as of 2016	Library gives IAP a WoS extract of current 2016 metrics for all existing 2004-2013 publications → IAP runs an Extract-Transform-Load (ETL) process to update all publication metrics in IAP SQL Server database
3B: Add a year of new publications (2014) to the existing 2004-2013 data to create a 2004-2014 dataset	Library gives IAP the Python output of new publications by author → IAP runs an ETL process that includes data integrity checks and produces publication reports by year



DATA PROBLEM #4

Problem	Manual process	Automated solution	Time savings
Calculating several bibliometric measures based on final publication dataset	Calculated using Excel formulas	SQL Server database automatically calculated ~200 bibliometric data points for all 6 research areas	Significant! ~50 hours



#watitis2016



UNIVERSITY OF
WATERLOO

AUTOMATED REPORTS

	A	B	C	D	E	Q	R
1	Overall metrics as of data year (2016)						
2							
3		Research area A	Research area B	Research area C	Research area D	Research area E	Research area F
4	data_year	2016	2016	2016	2016	2016	2016
5	H_index	42	79	30	43	30	86
6	Overall_Publications	939	2550	229	495	229	4276
7	Overall_Citations	8565	39750	3267	9190	3267	57294
8	Overall_Avg_Citations_per_Pub	9	16	14	19	14	13
9	Overall_Median_Citations	3	6	7	7	7	4
10	All_Pubs_No_Citations	301	480	32	61	32	1119
11	Overall_Pct_No_Citations	32.1	18.8	14	12.3	14	26.2
12	All_Pubs_Min_One_Citation	638	2070	197	434	197	3157
13	Overall_Pct_Min_One_Citation	67.9	81.2	86	87.7	86	73.8
14	All_Pubs_Min_Hundred_Citations	4	58	4	14	4	71
15	Overall_Pct_Min_Hundred_Citations	0.4	2.3	1.7	2.8	1.7	1.7
16							
17	Metrics by research group as of data year (2016)						
18							
19	RESEARCH_GROUP	DATA_YEAR	Avg_Avg_Pubs_per_Researcher				
20	Research area A	2016	5				
21	Research area B	2016	3				
22	Research area C	2016	3				
23	Research area D	2016	1				
24	Research area E	2016	4				
25							

Metrics
Research area A
Research area B
Research area C
Research area D
Research area E
Research area F



#watitis2016



UNIVERSITY OF
WATERLOO

IMPACT



- Python script and IAP automation processes saved 200+ hours of manual work

- Replicable
- Less prone to error
- High quality, robust data



#watitis2016



UNIVERSITY OF
WATERLOO

DOCUMENTATION

- Significant spot checking
- Data integrity analysis
- Process documentation
- 29 data dictionaries

Data dictionary criteria	
Project Title	
Metric Title	
IAP Technical Name	
Technical Definition	
Description	
Purpose	
Base Time	
Category	
Details	Inclusions
	Exclusions
	Time Frame
	Date Available
	Source
Owner	Performance Owner
	Primary Data Owner
	Primary Data Owner Contact
	Secondary Data Owner
Reported In	
Notes	



NEXT STEPS



Ongoing tweaking of Python script

Web of Science & InCites APIs

Online database to enhance vetting experience

Bibliometric data warehouse

Application to other discipline-level analyses



#watitis2016



UNIVERSITY OF
WATERLOO

BARRIERS

- Identifying expertise on campus
- Finding capacity to support automation needs



#watitis2016



UNIVERSITY OF
WATERLOO

Thank you!



jana.carson@uwaterloo.ca
shannon.gordon@uwaterloo.ca



#watitis2016



UNIVERSITY OF
WATERLOO

WORKS CITED

- University of Waterloo Working Group on Bibliometrics. (Winter 2016). *White paper: Measuring research output through bibliometrics*. Retrieved from https://uwspace.uwaterloo.ca/bitstream/handle/10012/10323/Bibliometrics%20White%20Paper%202016%20Final_March2016.pdf?sequence=4&isAllowed=y

