National Institute of Environmental Health Sciences
Your Environment. Your Health.

Tracking Research Impacts – Automated Research Impact Assessment

Christie Drew, Ph.D.
Program Analysis Branch, NIEHS
September 4, 2014

National Institute of Environmental Health Sciences
Your Environment. Your Health.

Tracking Research Impacts –

BEAT: Bibliography Extraction and Annotation Tool

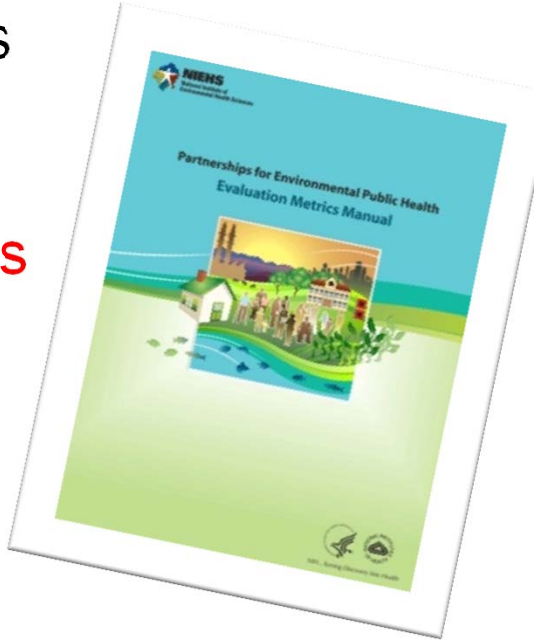
Christie Drew, Ph.D.
Program Analysis Branch, NIEHS
September 4, 2014

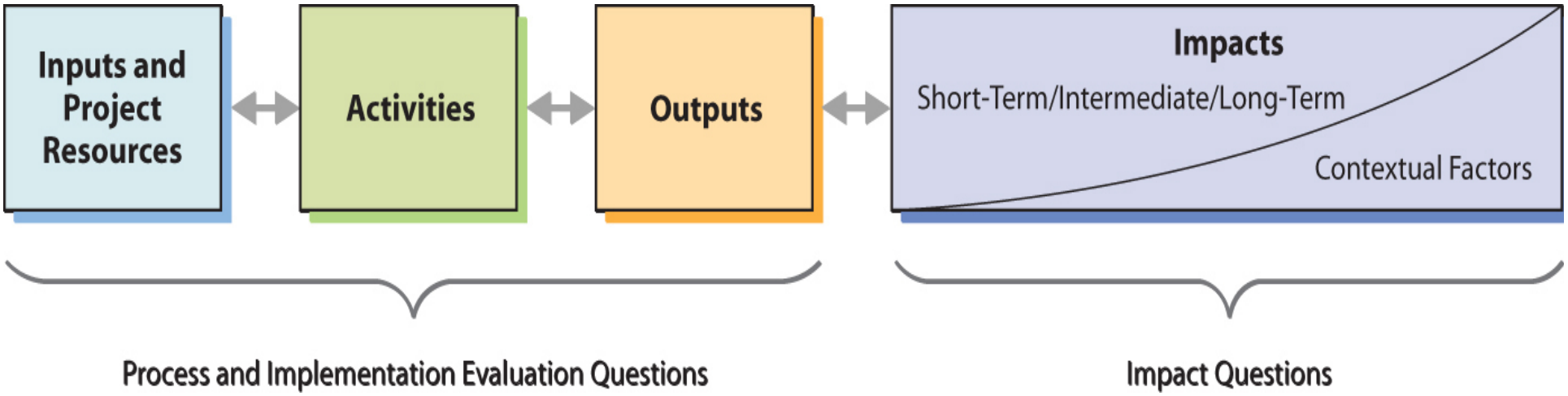
National Institute of Environmental Health Sciences

Mission: Reduce the burden of human illness and disability by understanding *how the environment influences the development and progression of human disease.*

Evaluation Context at NIEHS

- Constant questions about our portfolios
  - Topics, Methods, Approaches, Results, Impacts
- Logic modeling to go beyond bibliometrics
  - Publication content – not just count!
- Growing emphasis on outputs and impacts
  - Evaluation Metrics Manual
  - CareerTrac





Logic Model – organized, project specific, informs metrics

- **Inputs** – resources available
- **Activities** – actions that use available resources
- **Outputs** – direct products of activities
- **Impacts** – benefits or changes resulting from activities, outputs

ARIA Development

- Timeline:
  - 2012: Idea development; feasibility testing
  - 2013: Case study development; Peer reviewed paper development
  - 2014: Implementation in SPIRES
- Team (to date):
  - NIEHS PAB: Christie Drew, Kristi Pettibone
  - Open Intelligence: F.O. Finch; Doug Giles
  - NIH OER ORIS: Paul Jordan

Thought process/hypothesis

- Technology exists at NIH (SPIRES) to automate analysis of funding sources associated with a list of references
- Bibliography of an “important artifact” is an untapped resource for assessing impacts
- “Important artifact” = document from a credible source that is plausibly connected to NIEHS/NIH research
- Artifacts include:
  - Documentation of policy/regulatory decisions
  - Clinical and treatment guidelines
  - Major decision or guidance documents
  - Reference works from authoritative sources

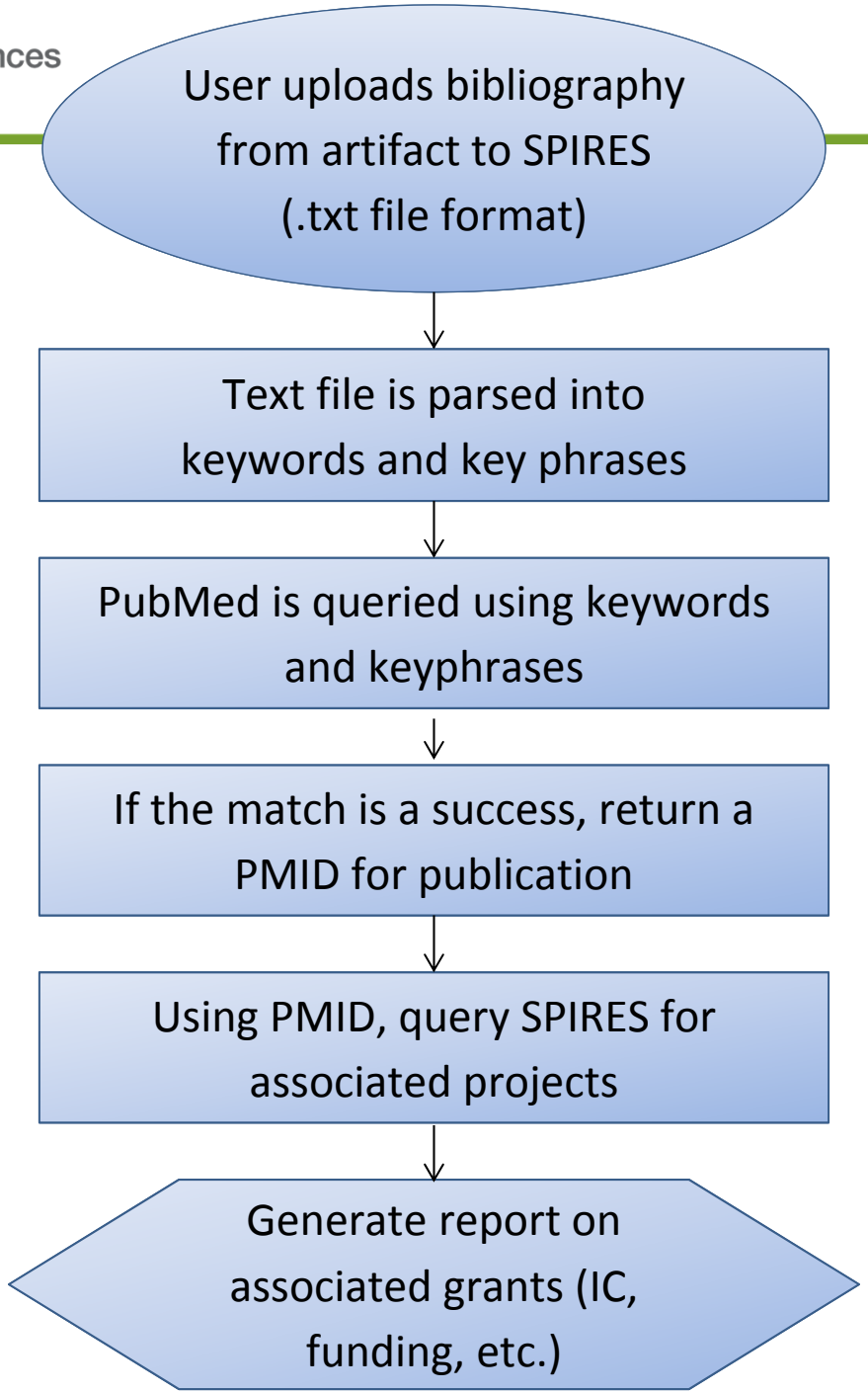
ARIA Process

Built within SPIRES tool

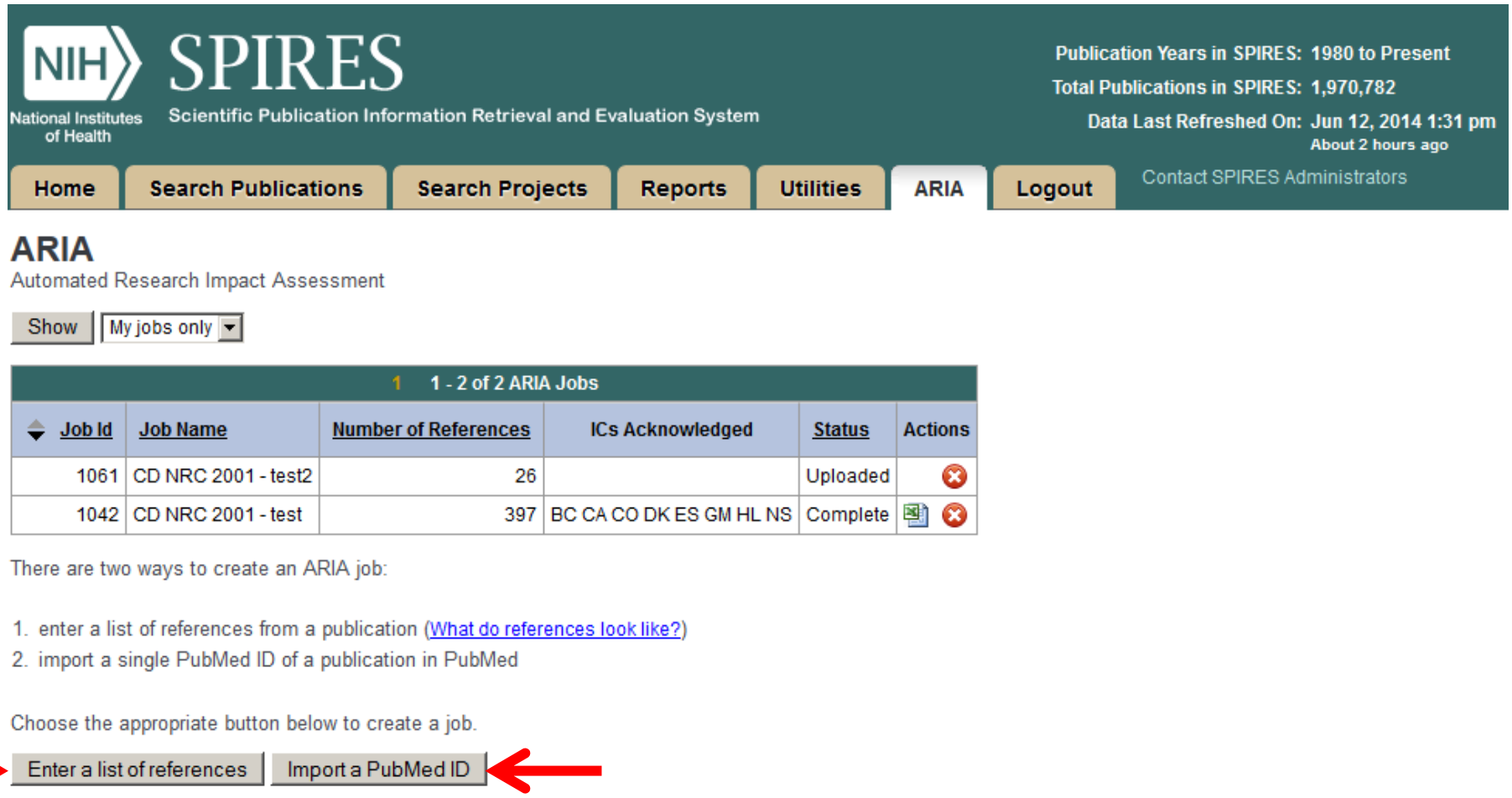
Imports any list of references (*.txt)

Creates new statistics

Analyses & summarizes results



ARIA Page in SPIRES



1. Paste references from .txt file

1001	Ambient particulate matter air pollution in Mpererwe District, Kampala, Uganda: a pilot study.	6	AI ES	Complete	

There are two ways to create an ARIA job:

1. enter a list of references from a publication (What do references look like?)
2. import a single PubMed ID of a publication in PubMed

Choose the appropriate button below to create a job.

Hide list form | Import a PubMed ID

Job Title: CD NRC 2001 - test2

Email: christina.drew @mail.nih.gov

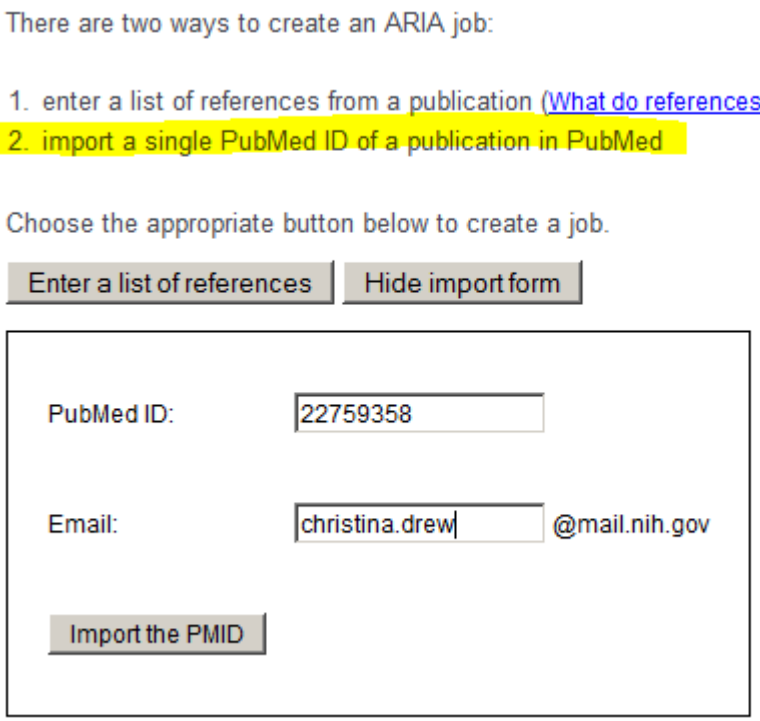
References to upload: (one reference per line)

Aposhian, H.V., E.S.Gurzau, X.C.Le, A.Gurzau, S.M.Healy, X.Lu, M.Ma, L.Yip, R.A.Zakharyan
Barchowsky, A., L.R.Klei, E.J.Dudek, H.M.Swartz, and P.E.James. 1999b. Stimulation of
Chen, N.-Y., W.Y.Ma, C.Huang, M.Ding, and Z.Dong. 2000a. Activation of PKC is required
De Kimpe, J., R.Cornelis, L.Mees, R.Vanholder, and G.Verhoeven. 1999b. 74As-arsena
De Kimpe, J., R.Cornelis, and R.Vanholder. 1999a. In vitro methylation of arsenite by ra
EPA (U.S. Environmental Protection Agency). 2000a. 40 CFR Parts 141 and 142. Natior
EPA (U.S. Environmental Protection Agency). 2000b. 40 CFR Parts 141 and 142. Natior
EPA (U.S. Environmental Protection Agency). 2000b. Estimated Per Capita Water Inges
EPA (U.S. Environmental Protection Agency). 2000c. Arsenic Proposed Drinking Water
EPA (U.S. Environmental Protection Agency). 2000c. Arsenic Proposed Drinking Water

Upload the References

1. Provide Job Title
2. Enter Email
3. Add references (1 per line)
4. Upload

2. Link to a Bibliography for an Artifact in PubMed



1. Provide Job Title
2. Enter Email
3. Add references (1 per line)
4. Upload

Status codes keep you up to date

ARIA
Automated Research Impact Assessment

References have been uploaded

Show | All jobs

1 - 19 of 19 ARIA Jobs

Job Id	Job Name	Number of References	ICs Acknowledged	Status	Actions
1061	CD NRC 2001 - test2	26		Uploaded	
1060	Practice parameter update: management issues for women with epilepsy--focus on pregnancy (an evidence-based review): vitamin K, folic acid, blood levels, and breastfeeding: report of the Quality Standards Subcommittee and Therapeutics and Technology A ...	38		Processing	
1059	Standards of medical care in diabetes--2011.	379		Processing	
1057	Ca2+-dependent, stimulus-specific modulation of the plasma membrane Ca2+ pump in hippocampal neurons.	43	AG DA DC EY GM HL MH NS	Complete	
1056	Integrated Science Assessment for Particulate Matter	4,686	AA AG AI AR AT CA CL CO CP DA DC DK DP EB ES GM HC HD HG HL HR MD NS OH RR TW	Complete	

Click the XLS icon ( ) to retrieve your file

Delete jobs using this icon

National Institute of Environmental Health Sciences
Your Environment. Your Health.

Raw Data Output (Project Mappings)

Title/Author/ Year Found	Published since 1980	PMID Found	Analyzed by ARIA	Why not?	Parsed Title	Parsed Authors	Parsed Pub Year	PMID	Confirmed Projects	Unconfirmed Projects	Original Reference Text
Yes	Yes	Yes	Yes		Subchronic dispositional and toxicological effects of arsenate administered in drinking water to mice.	M F Hughes; D J Thompson	1996	8874535			Hughes, M.F., and D.J.T Subchronic disposition: effects of arsenate adm water to mice. J. Toxicol 49(2):177-196.
No			No	Gray lit	Understanding Risk: Informing Decisions in a Democratic Society.		1996				NRC (National Researc Understanding Risk: Int Democratic Society. Wa Academy Press.
Yes	Yes	No	No	Unknown	Evidence of an immunologic mechanism behind the therapeutic effects of arsenic trioxide (As2O3) on myeloma cells.	S Deaglio; G Baj D Canella; S Waxman A Arnulfo; F Malavasi	2001				Deaglio, S., D.Canella, S.Waxman, and F.Malav an immunologic mecha therapeutic effects of ar on myeloma cells. Leuk
Yes	Yes	Yes	Yes		DMPS-arsenic challenge test. I. Increased urinary excretion of monomethylarsonic acid in humans given dimercaptopropane sulfonate.	H V Aposhian; M E Cebrian A Arroyo; L M Del Razo; R C Dart K M Hurlbut; H Kreppel D Gonzalez-Ramirez; A Smith H Speisky; P Ostrosky-Wegman M E Gonsebatt; M M Aposhian	1997	9223554	P30ES006694,P 42ES004940		Aposhian, H.V., A.Arroyc Razo, K.M.Hurlbut, R.C. Ramirez, H.Kreppel, H.S M.E.Gonsebatt, P. Ostrc M.M.Aposhian. 1997. DI test. I. Increased urinary monomethylarsonic aci dimercaptopropane sul Exp. Ther. 282(1):192-2
Yes	Yes	Yes	Yes		Relationship of urinary arsenic to intake estimates and a biomarker of effect, bladder cell micronuclei.	M L Biggs; L E Moore D A Kalman; M T Smith C Hopenhayn-Rich; A H Smith	1997	9219557	P42ES004705,P 30ES001896		Biggs, M.L., D.A.Kalmar C.Hopenhayn-Rich, M.T 1997. Relationship of u estimates and a bioma cell micronuclei. Mutat.
Yes	Yes	No	No	Book	Profile of urinary arsenic metabolites in children chronically exposed to inorganic arsenic	Del Razo; G G García-Vargas L M; Gómez-Muñoz M C Hernández; M E Cebrián	1999				Del Razo, L.M., G.G.Gar M.C.Hernández, Gómez Cebrián. 1999. Profile metabolites in children inorganic arsenic in Me Arsenic Exposure and H W.R.Chappell, C.O.Abe R.L.Calderon, eds. Oxfc
					The causes of cancer: quantitative estimates of						Doll, R., and R.Peto. 19 cancer: quantitative esti

Statistics

- Total # Submitted
- Total # Not Analyzed
  - Title, author or year not determined
  - Published before 1980
  - PMID not determined
- Total # that can be analyzed
- Total # of references that acknowledge NIH Grant
- Total # of references that acknowledge an ES Grant
- % of references that acknowledge NIH funding
- % of references that acknowledge ES funding
- % of NIH references from ES
- Total # NIH Grants referenced
- Total number of ES Grants referenced

Summary Output

ARIA Metrics		Artifacts
Total # of references submitted		625
Total # of references that could not be analyzed		238
Title, author or publication year could not be determined	31	
PMID could not be determined	198	
Published before 1980	9	
Total # of references that are analyzable		387
Total # of references that acknowledge an NIH Grant		12
Total # of references that acknowledge an ES Grant		11
% of references that acknowledge NIH funding	(12/387)	3%
% of references that acknowledge ES funding	(11/387)	3%
% of NIH references from ES	(11/12)	92%
Total number of NIH Grants referenced		14
Total number of ES Grants referenced		12

ARIA Metrics

	Artifacts		
	2009 EPA Particulate Matter ISA	2010 EPA Carbon Monoxide ISA	2012 EPA Lead (Pb) ISA
Total # of references submitted	3,483	179	625
Total # of references that could not be analyzed	1,517	28	238
Title, author or year not be determined	2	0	31
PMID could not be determined	1,502	24	198
Published before 1980	13	4	9
Total # of references that are analyzable	1,966	151	387
Total # of references that acknowledge an NIH Grant	467	58	12
Total # of references that acknowledge an ES Grant	357	16	11
% of references that acknowledge NIH funding	(467/1966) 24%	(58/151) 38%	(12/387) 3%
% of references that acknowledge ES funding	(357/1966) 18%	(16/151) 11%	(11/387) 3%
% of NIH references from ES	(357/467) 76%	(16/58) 28%	(11/12) 92%

Observations & Questions

- Wide range of references supported by NIH: 467, 58, 12
- Wide range of the proportion of NIH supported refs are from NIEHS: 357, 16, 11
- What does it mean?
- Is there a critical mass of references that are needed in order to have a credible analysis?
- Is there a discernible pattern in terms of which studies were cited?

Strengths

- Automation
- Limited opportunity for bias
- Ability to examine long-term impacts
- Makes use of existing, readily available information sources
- Relatively simple to implement
- Hopefully can be available to all of NIH

Limitations

- Not all artifacts have a bibliography (laws, policies)
- Improperly sourced references (getting better with recent NIH requirements)
- Not all journals included in PubMed
- Reference might not support the findings (e.g. retraction/rebuttals)
- Parser imperfect

NRC 2001: Analyzed by ARIA = "No"		Total	Percent of total
Total not analyzed		**129**	100%
Reasonable	Abstract	1	1%
	Book	13	10%
	Gray lit	38	29%
	In Chinese	1	1%
	Thesis	2	2%
	Total	**55**	**43%**
Planned	Pub date <1980	14	11%
	Total	**14**	**11%**
Needs work	Authors in title	27	21%
	Unknown	33	26%
	Total	**60**	**47%**

Enhancements

- Early versions didn't handle "2001a" well – recent enhancement fixed this – letters now stripped from years
- Duplicates are not filtered out
- Envisioning an iterative process… look at results, clean data, resubmit, improve results
  - What is causing authors to be parsed into the title field?
  - Can we replace a line entry with a PMID that we find manually?

More thoughts/ Next steps?

- On balance we think the benefits outweigh the limitations
- Adds to our tool box of analytical approaches
- Invites a lot of questions
- Needs vetting and interaction to determine true utility – *Alpha-test available on request*
- Are there other data sets that provide grant number linkages that could be supported? (E.g. high throughput screening/genetic/epigenetic)
- Meta analysis to understand ‘benchmarking’

Acknowledgements

- Kristi Pettibone,
  NIEHS
- F.O. Finch, Doug Giles
  Open Intelligence
- Paul Jordan
  National Institutes of Health
- Jim Onken
  National Institutes of Health
- Jim Corrigan
  National Cancer Institute

Questions?

Christie Drew
drewc@niehs.nih.gov
919-541-3319