Automated Research Impact Assessment (ARIA)

Christina H. Drew[1], Kristianna G. Pettibone[1], Fallis Owen Finch, III[2], Douglas Giles[2], Paul Jordan[3]

[1]National Institute of Environmental Health Sciences, Program Analysis Branch; [2]Open Intelligence, Inc.
[3]National Institutes of Health, National Institutes of Health, Office of Data Analysis Tools and Systems (OD OER)

Abstract

As federal programs are held more accountable for their research investments, The National Institute of Environmental Health Sciences (NIEHS) has developed a new method to quantify the impact of our funded research on the scientific and broader communities.

A pilot version of the assessment tool was developed for NIEHS. Ideally the tool will become available to all NIH Extramural Staff. ARIA includes new statistics that science managers can use to benchmark contributions to research by funding source. This new method provides the ability to conduct automated impact analyses of federal research that can be incorporated in program evaluations.

We apply ARIA to several case studies to examine the impact of NIEHS funded research, propose a number of questions that the new method raises, and discuss strengths and weaknesses of the approach.

On balance, we believe that the strengths outweigh the limitations and that ARIA represents another tool that NIH can use to describe impacts of its research investments.

Evaluation Context at NIEHS

- We get many questions about portfolios:
  - **About:** methods, approaches, results, impacts
  - **From:** program officers, Extramural Division leadership, NIEHS leadership, NIH, HHS, reporters, external stakeholders, etc.

- Logic models help us look beyond simple output metrics to think about long term impacts.[1-3]

Process and Implementation Evaluation Questions   Impact Questions
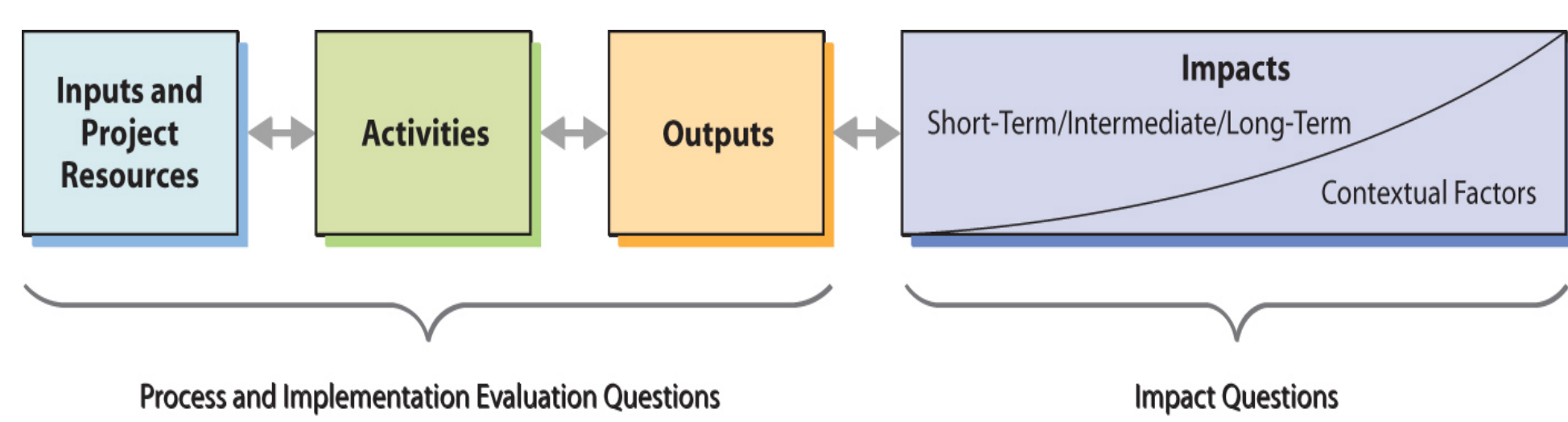
Logic Model — organized, project specific, informs metrics

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

- Typically evaluations start with NIH grant programs and look prospectively for impact.

- This tool provides an automated way to start with programs we know have had high impact and look retrospectively for NIH influence.
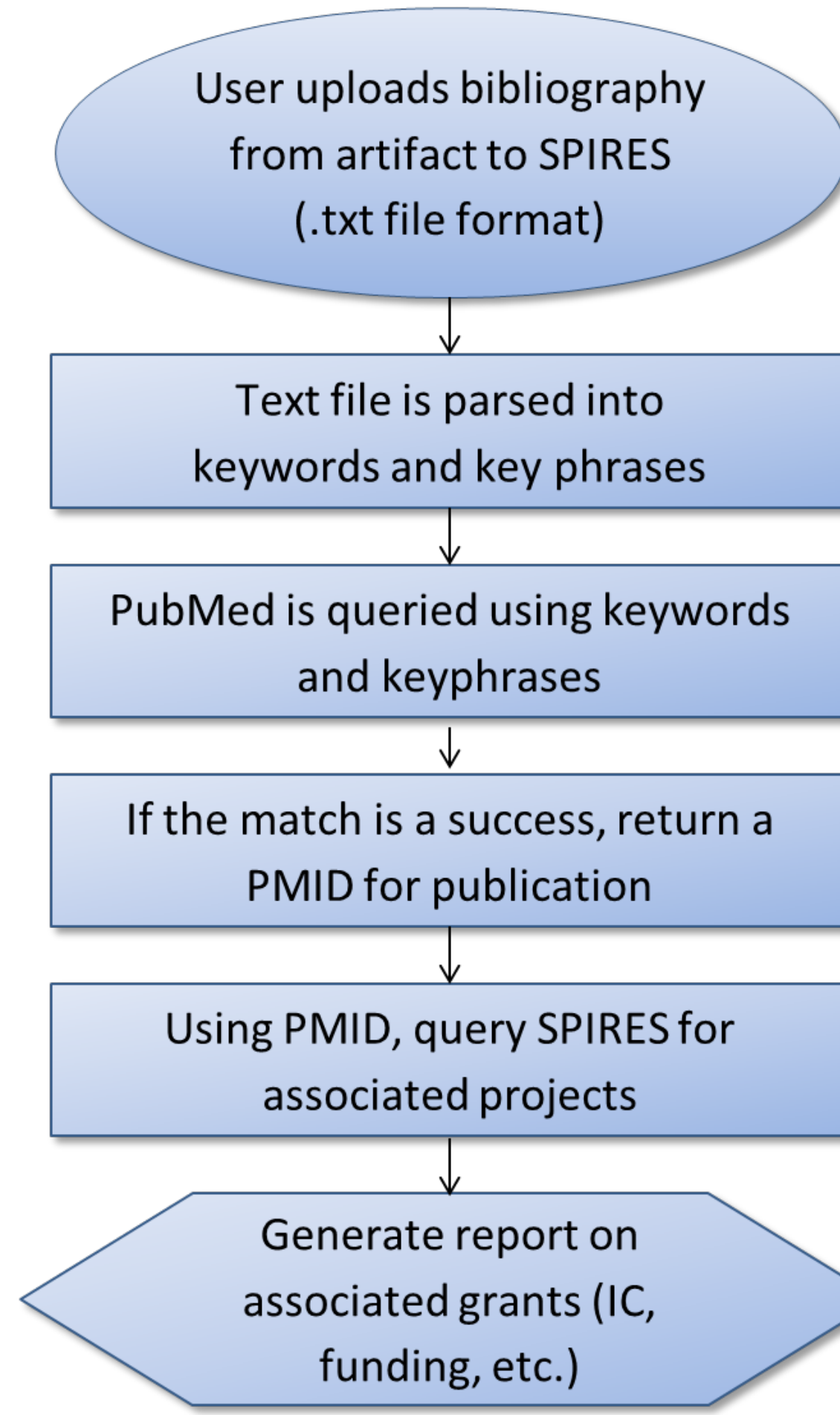
Premise

- Technology exists at NIH (SPIRES) to automate analysis of funding sources associated with a list of references
  - Scientific Publication Information Retrieval & Evaluation System[4]
  - Crawls PubMed and matches to NIH Grants
  - Provides information to QVR, RePORTER and has its own UI

- Bibliography of an "important artifact" is an untapped resource for assessing impacts
  - "Important artifact" = a 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 (Pilot)

User Actions:

1. Access ARIA Tool
2. Select "Enter list of References"
3. Provide Job Title
4. Enter Email
5. Add references (1 per line)
6. Hit upload button
7. Results load in job grid – status column indicates progress
8. Download file

In the Background:

User uploads bibliography from artifact to SPIRES (.txt file format)

1. Imports list of references

Text file is parsed into keywords and key phrases

2. Extracts title, author, and year from original reference into separate fields

PubMed is queried using keywords and keyphrases

3. Searches title, author, and year in PubMed and looks for PMID
Three separate parsers used to match with PubMed. Best results used.

If the match is a success, return a PMID for publication

Using PMID, query SPIRES for associated projects

4. If PMID found, looks for NIH Grant #

Generate report on associated grants (IC, funding, etc.)

5. Generates multi-tab MS Excel report with raw data and novel statistics about NIH project support

Raw Data Output

'Project Mappings' tab from the MS Excel output

Title/Author/ Year Found	Published since 1980	PMID Found	Analyzed by ARIA	Parsed Title	Parsed Authors	Parsed Pub Year	PMID	Confirmed Projects	Unconfirmed Projects	Original Reference Text
Yes	No		No	Effects and dose-response relationships of skin cancer and blackfoot disease with arsenic	Tseng WP	1977				Tseng, W.P. 1977. Effects relationships of skin cance disease with arsenic. Env 19 109-119.
Yes	No		No	Comparative study of chronic hydroarsenicism in two rural communities in the Laguna region of Mexico.	A Albores; I Tellez M E Cebrian; B Valdez	1979				Albores, A., M E Cebrian, B.Valdez. 1979. Comparat hydroarsenicism in two rur Laguna region of Mexico. Oficina Sant. Panam. 86(
Yes	Yes	Yes	Yes	Arsenic metabolites in hair, blood and urine in workers exposed to arsenic trioxide.	Y Yamamura; H Yamauchi	1980	7251401			Yamamura, Y., and H.Yam metabolites in hair, blood a exposed to arsenic trioxid 18(4) 203-210.
Yes	Yes	No	No	The causes of cancer: quantitative estimates of avoidable risk of cancer in the United States today.	R Doll; R Peto	1981				Doll, R., and R.Peto. 1981 cancer: quantitative estim cancer in the United State Cancer Inst. 66(6) 1191- 1
Yes	Yes	Yes	Yes	GSH release in bile as influenced by arsenite.	I Anundi; J Högberg; M Valter	1982	7128826			Anundi, I., J Högberg, and release in bile as influence Lett. 145(2) 285-288.
Yes	Yes	Yes	Yes	Medicinal arsenic and internal malignancies.	J Cuzick; M Gillman S Evans; D A Price Evans	1982	6212076			Cuzick, J., S Evans, M Gil malignancies. Br. J. Canc NRC (National Research C Nutrition and Cancer. Was Academy Press.
Yes	Yes	Yes	Yes	Diet, Nutrition and Cancer.	NRC	1982	7162996			Southwick, J.W., A.E.Wes J.Whitley, and R.Isaacs. 1 Health Associated with Ar Water in Millard County, 1 064 Health Effects Rese
				Community Health Associated with Arsenic in Drinking Water in Millard,						Environmental Protection /

- Raw data designed so user can easily recalculate metrics
- Original reference provided in right column
- Indicates if key criteria are met and included in automated analysis
  - Title / author / year found
  - Published since 1980
  - PMID found
  - Analyzed by ARIA
- Shows exactly what the parsers search
- Provides PMID, Confirmed projects
- Lists potential project matches (not included in summary statistics)

ARIA's Novel Metrics of NIH Investment and Case Studies

Objective	Metrics
Evidence of NIH investment	Total # and % of references that acknowledge NIH Project
Evidence of ICO investment	Total # and % of references that acknowledge an ICO Project
Relative investment of ICO compared to the rest of NIH	% of NIH references from ICO
Distribution of investment across NIH and ICO projects	Total # NIH/ICO projects referenced

	Artifacts		
Summary Output	2009 EPA Particulate Matter ISA[5]	2010 EPA Carbon Monoxide ISA[6]	2012 EPA Lead (Pb) ISA[7]
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 NIEHS Grant	357	16	11
% of references that acknowledge NIH funding	(467/1966) 24%	(58/151) 38%	(12/387) 3%
% of references that acknowledge NIEHS funding	(357/1966) 18%	(16/151) 11%	(11/387) 3%
% of NIH references from ES	(357/467) 76%	(16/58) 28%	(11/12) 92%

ICO = NIH Institute, Center or Office

NIEHS = National Institute of Environmental Health Sciences

We examined references for three Integrated Science Assessments available electronically from the Environmental Protection Agency.

Criteria for "important artifacts"
- **Plausible** – NIEHS reasonably expected to influence the artifact
- **Credible** – Artifact published by a trustworthy source
- **Important** – makes a significant contribution to the field of environmental health science

Observations
- Wide range of references supported by NIH
- % NIEHS/NIH support also ranges widely
- Many references not "parsable"
  - More work needed on this, but matching to PubMed is good
  - When a reference is analyzed most likely "gray lit" or books

DISCUSSION

Questions:
- What does it mean?
- Is there a critical mass of references that are needed in order to have a credible analysis?
- Can we determine "benchmarks" for specific fields or types of artifacts?

Strengths:
- Automated – requires a fraction of the time needed for manual analysis
- Ability to examine long-term impacts
- Makes use of existing, readily available information sources
- Relatively simple to implement
- Could 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. For example, deeper analysis of one ARIA report[8] found that, of 129 references not analyzed by ARIA
  - 14 (11%) published before 1980
  - 55 (43%) were "reasonable" – books abstracts, gray literature, non-english, or a thesis and thus not likely to be in Pub Med.
  - 60 (47%) unknown errors

Future Directions

- Hoping to expand pilot to broaden access to all of NIH via SPIRES
- Metrics need vetting and discussion within NIH analysis community to assess utility and meaning of results
- Potential algorithm enhancements:
  - Filter out duplicates
  - Allow user to import a combination of references and PMIDs
  - Track iterations of requests
  - Improve parser capacity (e.g., a common error is to interpret authors as the title, preventing possible match to PubMed record)
  - We have already added a filter to the year so that letters (e.g. 2001a) are removed

References

1. Engel-Cox, J. A., B. Van Houten, et al. (2008). "Conceptual model of comprehensive research metrics for improved human health and environment." Environ Health Perspect 116(5): 583-592.
2. Liebow, E., J. Phelps, et al. (2009). "Toward the assessment of scientific and public health impacts of the National Institute of Environmental Health Sciences Extramural Asthma Research Program using available data." Environ Health Perspect 117(7): 1147-1154.
3. Orians, C., J. Abed, et al. (2009). "Scientific and Public Health Impacts of the NIEHS Extramural Asthma Research Program - Insights from Primary Data." Res Eval 18(5): 375-385.
4. Boyack, K. W. and P. Jordan (2011). "Metrics associated with NIH funding: a high-level view." Journal of the American Medical Informatics Association 18(4): 423-431.
5. EPA (2009). Integrated Science Assessment for Particulate Matter. N. C. f. E. Assessment. Research Triangle Park, NC.
6. EPA (2010). Integrated Science Assessment for Carbon Monoxide. N. C. f. E. Assessment. Research Triangle Park, NC.
7. EPA (2012). EPA Integrated Science Assessment for Lead. E. P. Agency. Research Triangle Park, NC, Environmental Protection Agency.
8. NRC (2001). Update NRC Arsenic in Drinking Water. Washington, D.C.

Acknowledgments