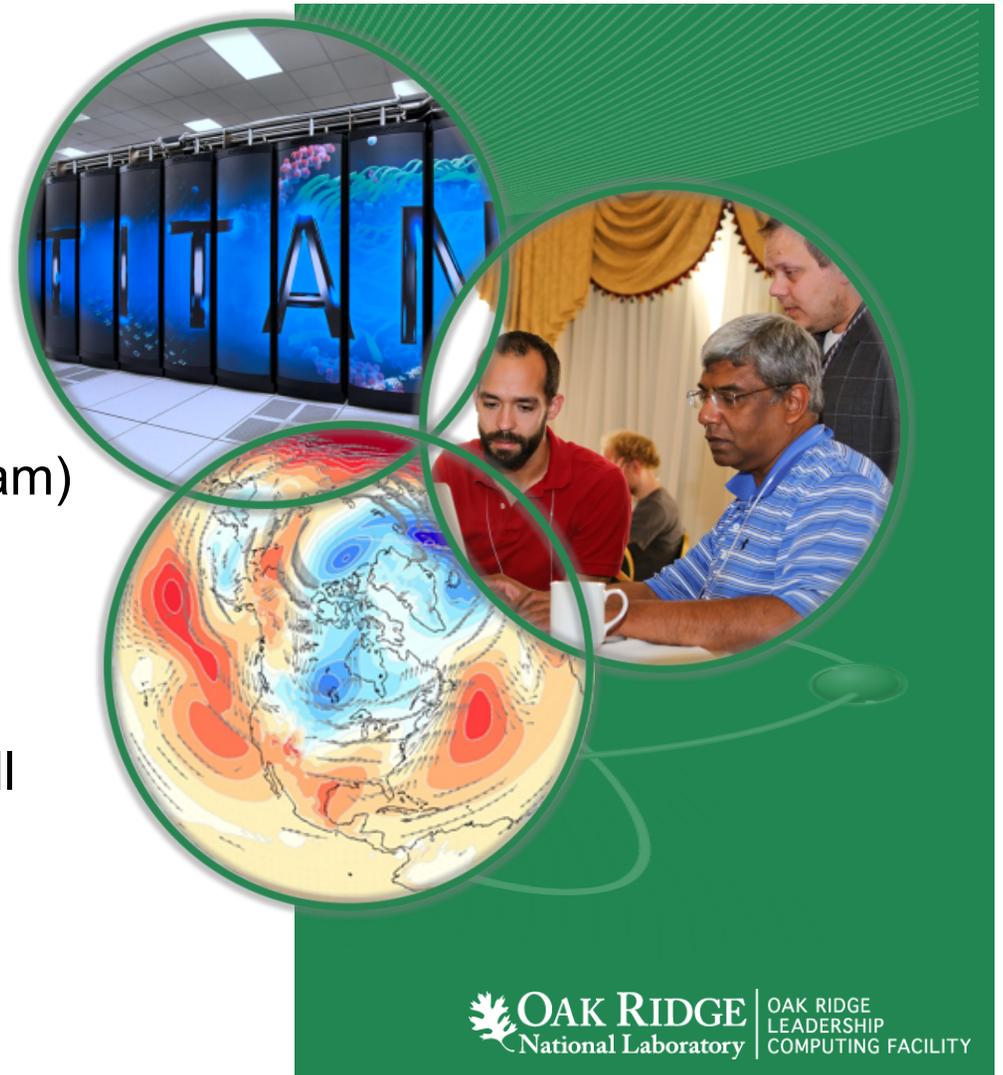


Discovery of Publications from Scientific User Facilities

Jack Wells
(for OLCF publications discovery team)
Oak Ridge National Laboratory

Presentation for
OLCF Monthly User Conference Call
June 29, 2016

ORNL is managed by UT-Battelle
for the US Department of Energy



Scientific publication curation

- Require an effective and efficient process to discover, verify and curate publications.
- No single data source or collection method, if implemented in isolation, is sufficient to the task, e.g.,
 - User-reported lists of publications are not always complete,
 - Automated searches can be no more complete than the completeness of the database searched.

**Common challenge for National Laboratories
and Academic Institutions**

DOE Office of Science, ASCR Facilities Division Operational Assessment Guidance: Section 3.1 Science Output (Publications)

The Facility tracks and reports the number of refereed publications written annually based on using (at least in part) the Facility's resources. For the LCFs, tracking is done for a period of five years following the project's use of the Facility. This number may include publications in press or accepted, but not submitted or in preparation.

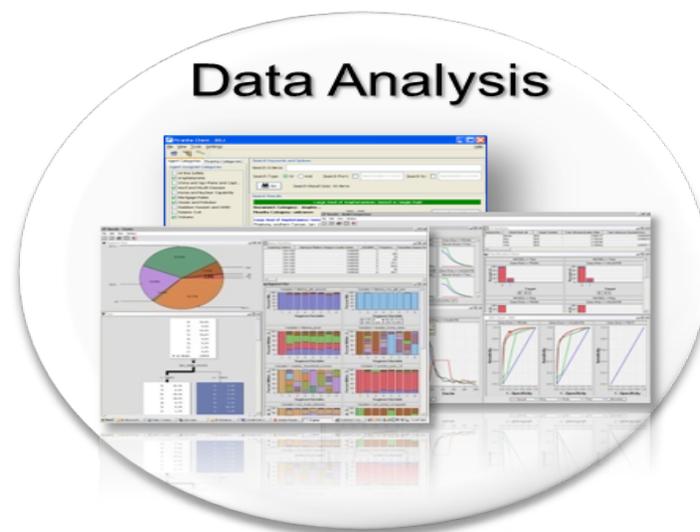
- From *CY15 DOE Office of Science, High-Performance Computing & Networking Facility Operational Assessment & Facility Metric Updates*
December 2015

Exploring new publication discovery tools

- Tools and procedures developed within ORNL data science research have advanced ability to track and report publication outcomes:
 - Discovery across multiple years,
 - Ability to refresh information more frequently, and
 - Find publications not reported through, e.g., users' project reports.
 - Early success: Discovery of *Nature* paper published in Sept. 2012 from 2009 INCITE project.
- Implemented process of curation, integrating automated database searches with user-reported publications creating a validated repository of publications for institutional use,
 - Well defined process provides high data quality.
 - Disambiguation of acknowledgements requires communication with users.
- Validated results can be communicated with confidence
 - Meeting reporting requirements of sponsor
 - More readily communicate outcomes to the full spectrum of stakeholders.

Providing a sustainable publications discovery service

- Deliverables
 - Validated facility publications database with regular updates (e.g., weekly)
 - Publications Dashboard displaying up-to-date key performance indicators
 - Maintenance of publication repository for operational purposes
 - For example, this capability enables regular citation-count updates.



Publications discovery

- Four data collection sources:
 - Automated Publication Discovery (ADP or “Cobra”)
 - Self-Reports (SR) from users reporting to the facility
 - PTS (Publication Tracking Service) for ORNL staff
 - Google Scholar Alerts (newest method – just added)
- Process:
 - Text Analytics: Intelligent Search for Key Words (e.g., User Name, Facility Name, and/or INCITE/ALCC Acknowledgement), and
 - Validated and Curated Results Integrated through online database
 - Given Publication DOI, Data Input from ISI WoS is automated
 - Manual integration of ADP-Cobra with SR and PTS
 - Automated scripts for duplicate detection
 - Scripts to Lookup Journal Impact Factor (JIF) and ISI Citation Count
 - Manual disambiguation of subset of acknowledgements

Data Sources: Strengths & Weaknesses

Data Source	Description	Strengths	Weaknesses
ADP COBRA	Automated search and retrieval tool utilizing structured API queries against proprietary bibliometric databases	<ul style="list-style-type: none"> • Data Quality • Accuracy • Automation • Retrieval Speed • Flexibility • Extensibility • Validation 	<ul style="list-style-type: none"> • Immediacy (latency of WoS) • Coverage*
Self Reports	Publications which are disclosed to user facility via reporting requirements per program or informal report	<ul style="list-style-type: none"> • Direct User Communication • Immediacy • Coverage • Metadata (Project ID) 	<ul style="list-style-type: none"> • Data Quality • Integration • Relevance • User Discipline • Validation
PTS (ORNL Staff)	Publications which are disclosed to Laboratory system per requirement for all ORNL R&D staff	<ul style="list-style-type: none"> • Immediacy • Coverage • User-Staff Joint Pubs 	<ul style="list-style-type: none"> • Data Quality • Integration** • User Behavior • Validation
Google Scholar	<i>Publication discovery utilizing alert features with limited keyword queries</i>	<ul style="list-style-type: none"> • <i>Immediacy</i> • <i>Coverage</i> • <i>Extensibility</i> 	<ul style="list-style-type: none"> • <i>Integration</i> • <i>Data Quality</i> • <i>Relevance</i>

We are refining our process and tools to improve effectiveness:

- Terminated less productive manual library searches
- Leverage automated method as primary discovery method
- Reduce staff effort recognizing support will always be needed
- Improving quality of user reports and PTS repository
- Daily Updates to confirmed database
- Monthly Updates to OLCF Webpage
 - <https://www.olcf.ornl.gov/leadership-science/publications/>

OLCF now reports high-quality data covering multiple years

CY	Unique Publications	High Impact ISI Journals (JIF>7)	High-Impact ISI Journals (JIF>8)
2016	144	22	15
2015	344	42	28
2014	293	42	21
2013	343	27	9
2012	315	41	22

Data sampled: 28 June 2016

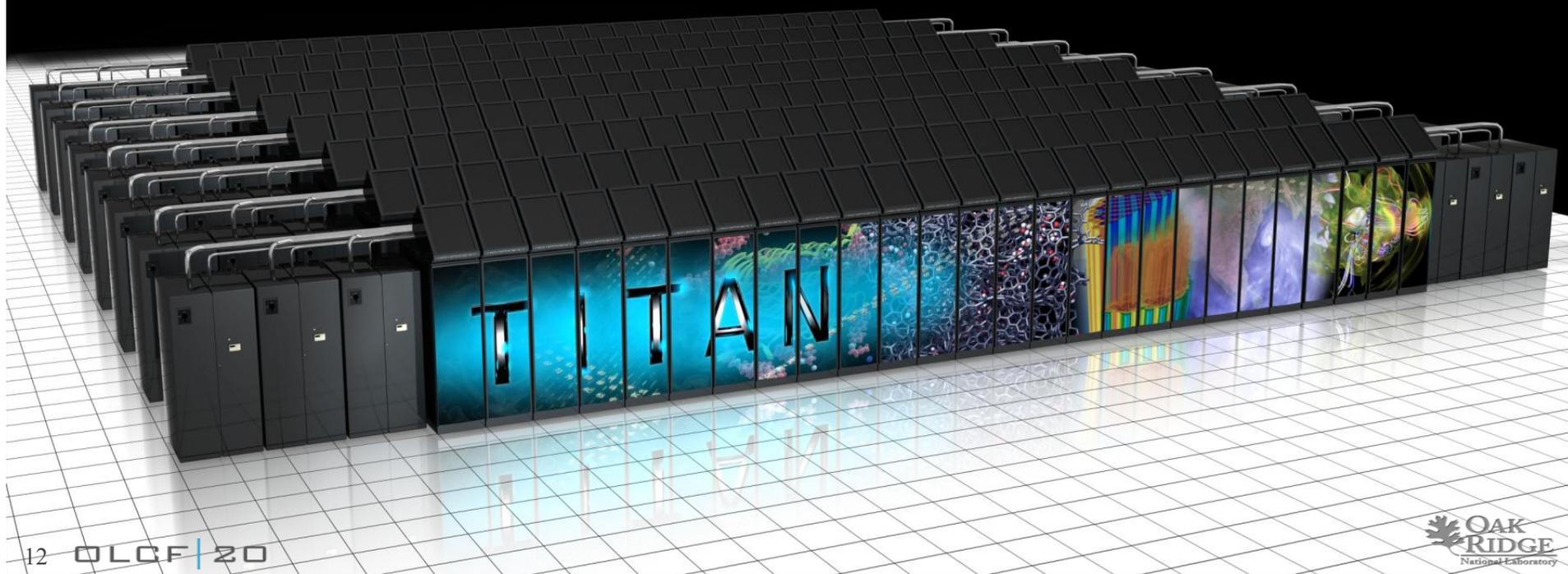
Continuous Improvement

- Automated discovery of publications improved by expanding search methodology and refinement of keywords
- Service process improved by established acceptance criteria, validation methodology and reporting
- Application of Lessons Learned to improve discovery methodology and overall service delivery
- Validated publication database enables communication of program outcomes
- Publication database will enable new analysis of publications

How can you help?

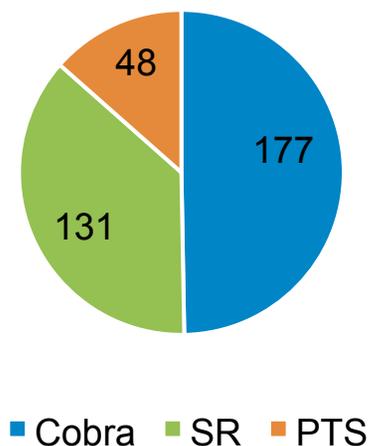
- Be sure to include the OLCF acknowledgement statement in all of your publications. The statement can be found on the OLCF website at <https://www.olcf.ornl.gov/media-center/media-kit/>.
- Be sure to report your publications in the quarterly and self-reports. If a publications occurs after your project is over, please notify the OLCF of the publication.
- Check to see if your publications are captured for 2015 and 2016 on the OLCF website at <https://www.olcf.ornl.gov/leadership-science/publications/>.
- Coming soon – users will be able to interact with the publication data directly.

Questions?

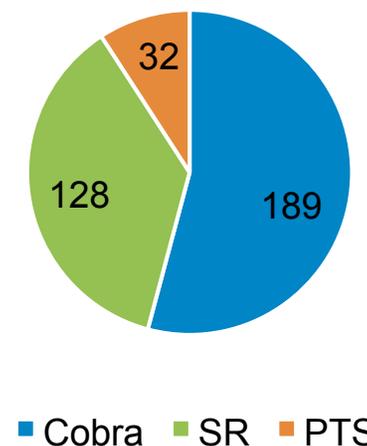


Cobra provides more valid records than the other two data sources

Number of Records Per Process
356 Valid Records in 2014

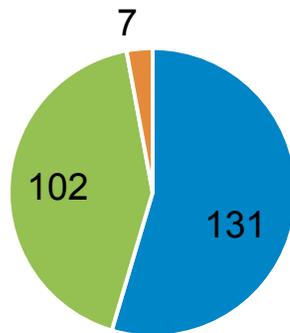


Number of Records Per Process
349 Valid Records in 2015



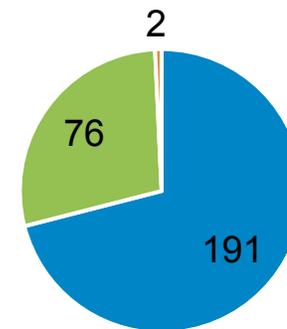
The majority of publications are identified by one and only one data source

240 Unique Publications in 2014
Number of Methods that identified a confirmed paper:



■ 1 Process ■ 2 Process ■ 3 Process

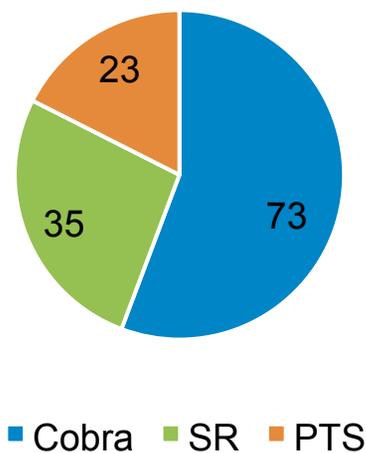
269 Unique Publications in 2015
Number of Methods that identified a confirmed paper:



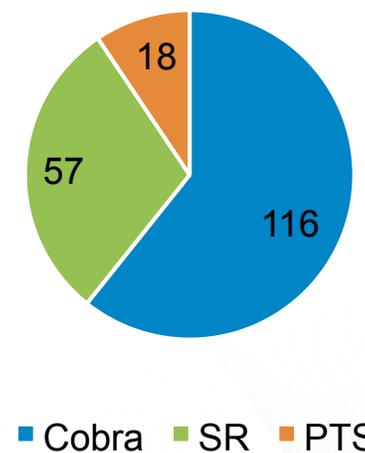
■ 1 Process ■ 2 Process ■ 3 Process

Each data source contributes to the unique publications

131 Publications Found by One and Only One Source, OLCF 2014



191 Publications Found by One and Only One Source, OLCF 2015

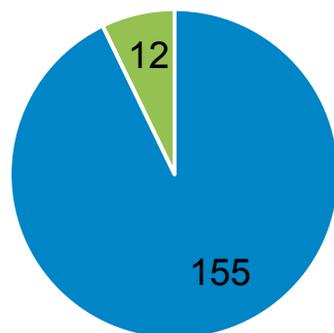


Data sampled: 21 January 2016

If the Cobra data source is removed from the curation process, how many publications would OLCF document for 2014 & 2015?

Number of unique publications without ADP Cobra:
2014 OLCF 167

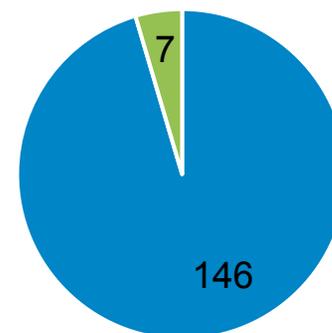
70% of
2014 Total



■ 1 Process ■ 2 Process

Number of unique publications without ADP Cobra:
2015 OLCF 153

57% of
2015 Total



■ 1 Process ■ 2 Process