

Comparative Methods of CCSG Data Gathering: Biosketches, Publications, Grants

A look at the Past , Present, and Future of Data Collection

CCAF-IT 2017
Ben Busby, Mahendra Yatawara, & Susan Sharpe

Biosketches: sometimes member data collection is like herding cats...

Susan Sharpe, MA

Biosketches: The expectation

- Routine process, simple 5 page CV of relevant work & interests.
- SciENcv – latest and greatest template.
- Expected Gathering Process: Ask and receive.

Biosketch Request
Sharpe, Susan C.

Dear CCSG Member,

In order to prepare for the 2015 CCSG Site Visit we are asking for all key personnel to check and send an updated copy of their NIH Biosketch to Susan Sharpe at the library by Friday April 1, 2015.

The Biosketch should be generated in the newest NIH format - preferably using the [SciENcv tool](#) from NCBI. If you should not need to use the tool - please refer to the attached NIH guidelines.

About the new format:

- Biosketches using the new format are limited to 5 pages instead of 4.
- The Contributions to Science section is replacing Detailed Peer-Reviewed Publications.
- Each Contribution can include up to four peer-reviewed publications.
- Applicants can include a link to list of their published work as found in a publicly available digital database such as My Bibliography.

About creating a personal statement:

- Personal Statements should mention your CCSG program alignment in addition to your scientific background and contributions. The statement can be something like

The reality: a process firmly rooted in the past

1. Email all required personnel request for Biosketch. Include links to SciENCv, provide Word Template, & latest instructions.
2. Wait. Some biosketches return. Edit. Store locally or send to shared drive.
3. Email personnel request for Biosketch Reminder. Add high priority message to email.
4. Wait. Some biosketches return. Edit. Store locally or send to shared drive.
5. Send messages to Faculty Leaders asking for support and encouragement.
6. Wait. Some biosketches return. Edit. Store locally or send to shared drive.
7. Rinse-Repeat x10 times.
8. Biosketches are gathered. Review and make final edits.

Search All Mail Items (CCH-E)		Received	Size	Cat...
Date Older				
0	Mail - RE: CCGS Site Visit Biosketch Request (2nd Notice) (Response Required) (Time Sensitive)!!	Thu 4/14/2016 2:23 PM	95 KB	
0	Mail - RE: Biosketch Request (URGENT!!) (RESPONSE REQUIRED!!)	Thu 4/13/2016 1:54 PM	23 KB	
0	Mail - FW: CCGS Site Visit Request for Biosketch (Time Sensitive) (Response Requested) (April 2016)	Mon 3/14/2016 2:22 PM	87 KB	
0	Urg - FW: CCGS Site Visit Request for Biosketch - Please respond by March 18th	Tue 3/1/2016 1:08 PM	87 KB	
0	Adm - NIH Biosketch (Response Requested) (Time Sensitive)	Fri 1/15/2016 12:24 PM	92 KB	
0	Adm - RE: Biosketch needed for CCGS Renewal (Time Sensitive)	Thu 7/15/2015 9:39 PM	25 KB	
0	Spill - RE: Biosketch needed for CCGS Renewal	Tue 7/14/2015 12:03 PM	20 KB	

Names redacted to protect the guilty.



SciENCv:

What went right:

- Automatically puts information in new format
- Create multiple versions
- Share entry and upkeep responsibilities with delegates
- Create sharable URL
- Links to MyBibliography

What went wrong:

- URL version doesn't enable viewers to download.
- De-centralized management (PI-centric, instead of institutionally)
- No delivery mechanism:
 - PDFs & Emails can be lost, forgotten, deleted, etc.



How do members of CCAF gather Biosketches?

- 69% of respondents rely on Members to submit and maintain Biosketches
- 14% have homegrown systems that centralize and keep track of Biosketches
- 4% have some sort of vendor system

Biosketch Methods of Collection:	# of Responses:
We rely on Members to submit and maintain Biosketches.	36
We use a homegrown system to collect, create, manage, & store.	7
Other: Members write, we edit or provide templates.	4
Homegrown Other: Yes, we have a homegrown solution, but...	1
Vendor Other: Yes, we have a vendor solution, but...	1
We use a vendor supported system to collect, create, manage, & store.	1
We use existing NIH provided tools (NCBI, etc).	1
Grand Total	49



Publications

How Moffitt collects Pub Data:

- Nightly search of author names via API to MEDLINE
- Download into holding queue
 - Impact Factor automatically assigned
- Daily author verification screening by human

View	Year	Month	PMID	Citation
VIEW	2008	Dec	18682882	Sanchez JA, Vogel JD, Kalish MF, Groner MP, Kado H, Lee JK, Medina K, Bink K, Burkhart Dec37(23):413-418. PubMedId: 18930709.
VIEW	2008	Dec	18930709	



What works for us, may not work for you:

Pros:

- Automated & customizable search algorithm
- Very little need for author input
- Standardized citation information
- Ability to pull corresponding data: Grant IDs, ORCID, MeSH, IF

Cons:

- Labor intensive
- Centralizing Screening process requires dedicated staff members
- Author Name Disambiguation remains a stumbling block



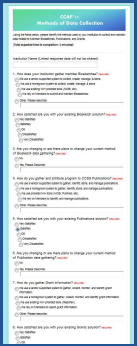
Publications: The current state and a look at our center's process

Mahendra Yatawara, MBA



CCAF-IT 2017 Survey


- <http://moffitt.libsurveys.com/CCAFData>
- Survey sent out April 20th
- Survey active until May 2nd
- Institutions responding: 44



NIH U.S. National Library of Medicine 13



Institutions Responding



NIH U.S. National Library of Medicine 14



How do Centers manage Pubs for CCSG?

Publications Systems

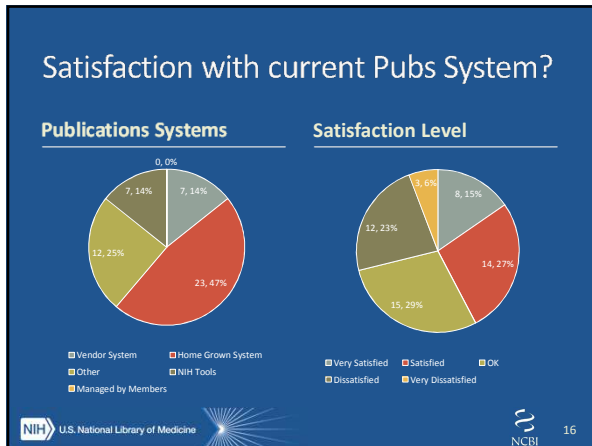
System	Percentage
Home Grown System	23, 47%
Other	12, 25%
Managed by Members	7, 14%
NIH Tools	7, 14%
Vendor System	0, 0%

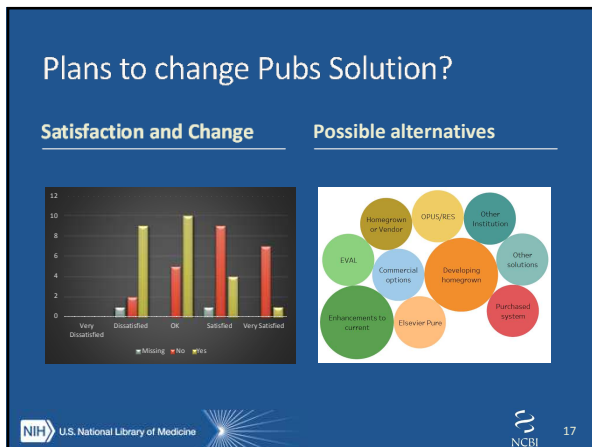
Vendor and Other

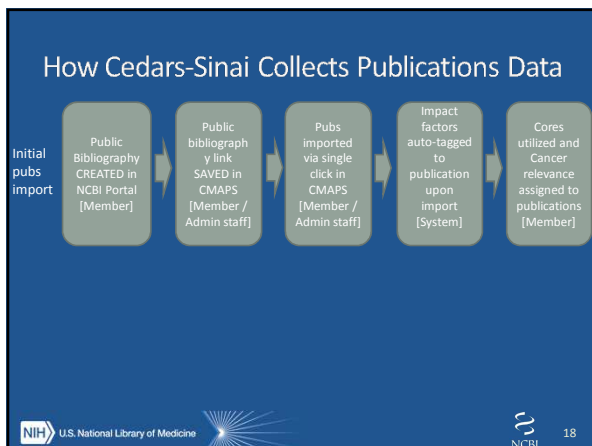
CAFÉ by USC	3
Opus/EVAL by Forte	2
Lattice Grid	2

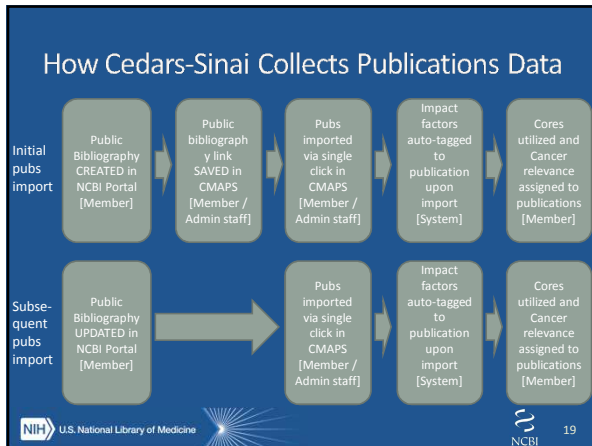
NIH U.S. National Library of Medicine 15











How is this process working for us?

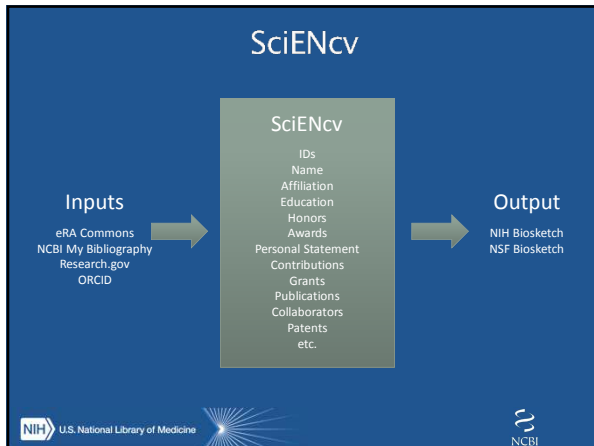
Key Advantages	Limitations
<ul style="list-style-type: none"> Reduction in non-value added work from CC Admin Members maintain in single location (NCBI portal) Auto-assignment of Impact Factor One-click reports 	<ul style="list-style-type: none"> Reminders for Members to keep NCBI Bibliography up-to-date Reminders for Members to allocate Core usage and Cancer Relevance to pubs in CMAPS

Logos for NIH U.S. National Library of Medicine and NCBI are present at the bottom.

The Futures: Biosketches, Grants, Pubs... and Data!

Ben Busby, NCBI

Logos for NIH U.S. National Library of Medicine, CEDARS-SINAI, MOFFITT CANCER CENTER, and NCBI are present at the bottom.



NCBI

My Bibliography

Search NCBI databases

Search: PubMed

My Bibliography

Recent Activity

Date	Database	Type	Term
01/17/16	Bioc	report	10.1002/anie.201511000
04/07/16	Bioc	report	10.1002/anie.201511000
27/Jan/2015	NCBI	search	10.1002/anie.201511000
27/Jan/2015	NCBI	search	10.1002/anie.201511000
27/Jan/2015	NCBI	search	10.1002/anie.201511000
27/Jan/2015	NCBI	search	10.1002/anie.201511000
27/Jan/2015	NCBI	search	10.1002/anie.201511000
27/Jan/2015	NCBI	search	10.1002/anie.201511000
27/Jan/2015	NCBI	search	10.1002/anie.201511000

Saved Searches

Collections

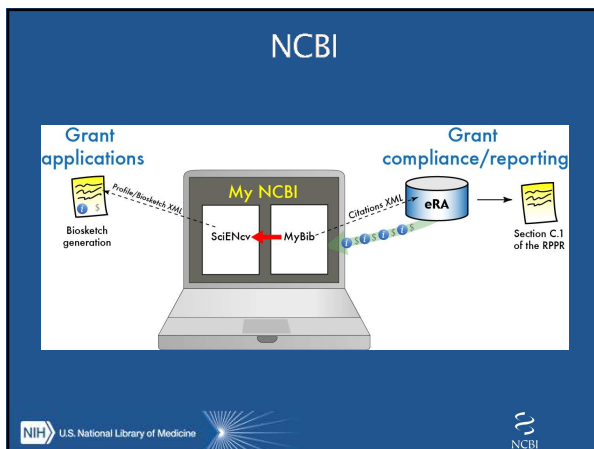
Collection Name	Items	Import/Export	Type
SciENcv	0	0	Standard
10.1002/anie	1	0	Standard
10.1002/anie	0	0	Standard

Filters

SciENcv

Field	Value	Priority	Type
ORCID iD	0000-0001-9152-0001	High	Profile
ORCID iD	0000-0001-9152-0001	High	Profile
ORCID iD	0000-0001-9152-0001	High	Profile

NIH U.S. National Library of Medicine NCBI



Better PubMed Searches!

U.S. National Library of Medicine
NCBI



For more information go to: ncbi.nlm.nih.gov/learn

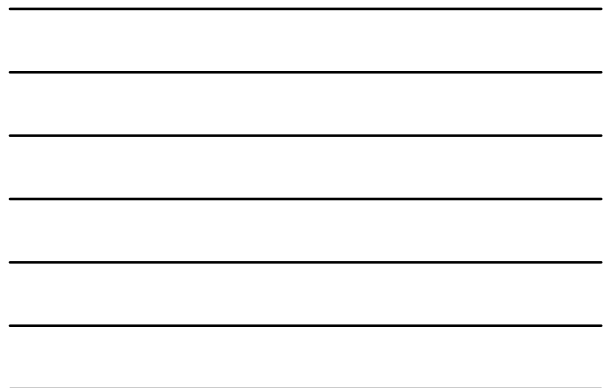
U.S. National Library of Medicine
NCBI



E-Utilities (Eutils)

Entrez Database	UID common name	E-utility Database Name
BioProject	BioProject ID	bioproject
BioSample	BioSample ID	biosample
BioSystems	BSID	biosystems
Books	Book ID	books
Conserved Domains	PFSSM-ID	cdcd
dbGAP	dbGAP ID	gap
dbVar	dbVar ID	dbvar
Epigenomics	epigenomics ID	epigenomics
EST	GI number	nucst
Gene	Gene ID	gene
Genome	Genome ID	genome
GEO Datasets	GDS ID	gds
GEO Profiles	GEO ID	geoprofiles
GSS	GI number	nucgss
HomoGene	HomoGene ID	homologene
MapR	MapR ID	mapr
NCBI C++ Toolkit	Toolkit ID	toolkit
NCBI Web Site	Web Site ID	ncbisearch
NLM Catalog	NLM Catalog ID	nlmcatalog
Nucleotide	GI number	nucleotide
PopSet	PopSet ID	popset
Probe	Probe ID	probe
Protein	GI number	protein
Protein Clusters	Protein Cluster ID	proteinclusters
PubChem BioAssay	AID	pcassay
PubChem Compound	CID	pcocompound
PubChem Substance	SID	pcsubstance
PubMed	PMID	pubmed
PubMed Central	PMCID	pmed
SNP	rs number	snp
SRA	SRA ID	sra
Structure	MMDB-ID	structure
Taxonomy	TaxID	taxonomy
UniGene	UniGene Cluster ID	unigene

U.S. National Library of Medicine
NCBI



Introducing... Entrez Direct The E-utilities on the UNIX command line

```


esearch -db gene -query "foxp2[gene]
AND human[orgn]" | \

elink -target protein -name
gene_protein_refseq | \

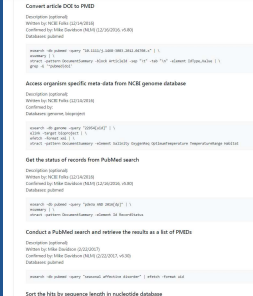
efetch -format fasta

```


ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/



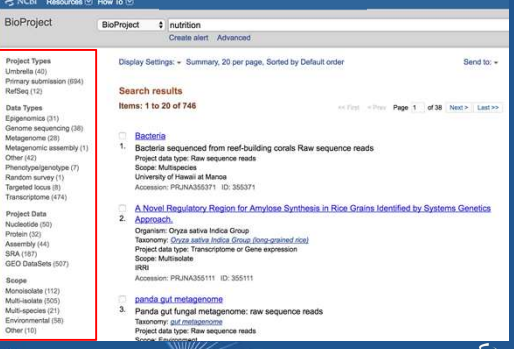
The EDirect Cookbook!



Google for EDirect Cookbook



BioProject



Project Types: 400
Primary submission (894)
RefSeq (12)


Data Types: 21
Ergonomics (2)
Genome sequencing (38)
Metagenome (2)
Metagenomic assembly (1)
Other (42)
Phenotypic/genotype (7)
Random survey (1)
Targeted locus (3)
Transcriptome (174)

Project Data: 505
Protein (32)
Assembly (44)
SRA (197)
GEO Datasets (507)

Scope: 112
Mono-locus (105)
Multi-species (21)
Environmental (38)
Other (13)

Search results: 1 to 20 of 746

- Bacteria**
1. Bacteria sequenced from reef-building corals Raw sequence reads
Project data type: Raw sequence reads
Scope: Multispecies
University of Hawaii at Manoa
Accession: PRJNA355371 ID: 355371
- A Novel Regulatory Region for Amylose Synthesis in Rice Grains Identified by Systems Genetics Approach**
Organism: *Oryza sativa* Indica Group
Taxonomy: *Oryza sativa* Indica Group [kingdom: eukaryota]
Project data type: Transcriptome or Gene expression
Scope: Multisite
RRR
Accession: PRJNA355111 ID: 355111
- parhva* sp. metagenome**
3. *Parhva* sp. fungal metagenome: raw sequence reads
Taxonomy: *parhva* sp.
Project data type: Raw sequence reads
Scope: *parhva* sp.



Minimizing Data Transfer

NCBI Resources How To

Nucleotide Nucleotide Advanced

FASTA +

Human endogenous retrovirus HERV-K, pol gene

GenBank: Y12391.1

Color key for alignment scores

<40 40-60 60-80 80-200 >=200

Query 1 450 900 1350 1800 2250

NIH U.S. National Library of Medicine NCBI

Minimizing Data Transfer

NCBI News Search NCBI

NCBI is currently testing https on public web servers until 4:00 PM EDT (20:00 UTC) today. You may experience problems with NCBI services during this test. [Read more.](#)

Introducing Magic-BLAST

Tuesday, September 22, 2016

Magic-BLAST is a new tool for mapping large sets of next-generation RNA or DNA sequencing runs against a whole genome or transcriptome. Magic-BLAST is available for UNIX, Mac OS X, and Windows as well as the source files are available on the [GitHub](#).

Each alignment optimizes a composite score, taking into account simultaneously the two reads of a pair, and in case of RNA-Seq, locating the candidate introns and adding up the score of all exons. Sequencing reads can be provided as NCBI SRA accessions, FASTA or SRA files.

Magic-BLAST implements ideas developed in the NCBI Magic queries using the NCBI BLAST libraries. Magic-BLAST is under active development, and we expect the next few releases to occur on a monthly basis. Read more about Magic-BLAST on the [EET site](#).

NIH U.S. National Library of Medicine NCBI

