Forthcoming: *Annals of Economics and Statistics (Annales d'Economie et Statistique)*, Issue 115/116, in press 2014

NBER WORKING PAPER SERIES

COMMUNITYWIDE DATABASE DESIGNS FOR TRACKING INNOVATION IMPACT: COMETS, STARS AND NANOBANK

Lynne G. Zucker
Michael R. Darby
Jason Fong

Working Paper No. 17404
http://www.nber.org/papers/w17404

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2011
Revised March 2014

The construction of Nanobank was supported under major grants from the National Science Foundation (SES-0304727 and SES-0531146) and the University of California's Industry-University Cooperative Research Program (PP9902, P00-04, P01-02, and P03-01). Additional support was received from the California NanoSystems Institute, Sun Microsystems, Inc., UCLA's International Institute, and from the UCLA Anderson School's Center for International Business Education and Research (CIBER) and the Harold Price Center for Entrepreneurial Studies. The COMETS database (also known as the Science and Technology Agents of Revolution or STARS database) is being constructed for public research use under major grants from the Ewing Marion Kauffman Foundation (2008-0028 and 2008-0031) and the Science of Science and Innovation Policy (SciSIP) Program at the National Science Foundation (grants SES-0830983 and SES-1158907) with support from other agencies. Our colleague Jonathan Furner of the UCLA Department of Information Studies played a leading role in developing the methodology for selecting records for Nanobank. We are indebted to our scientific and policy advisors Roy Doumani, James R. Heath, Evelyn Hu, Carlo Montemagno, Roger Noll, and Fraser Stoddart, and to our research team, especially Amarita Natt, Hsing-Hau Chen, Robert Liu, Hongyan Ma, Emre Uyar, and Stephanie Hwang Der. Additional support for COMETS has been provided by The Ewing Marion Kauffman Foundation which both hosts the main COMETS site and in the past ran a COMETS Travel Grants Program to support early use of the COMETS data and presentation of resulting empirical research at conferences through direct grants to users. Certain data included herein are derived from the High Impact Papers, Science Citation Index Expanded, U.S. State Indicators, and U.S. University Indicators of the Institute for Scientific Information®, Inc. (ISI®), Philadelphia, Pennsylvania, USA: © Copyright Institute for Scientific Information®, Inc. 2000-2003. All rights reserved. Certain data included herein are derived from the Nanobank and the Connecting Outcome Measures of Entrepreneurship, Technology, and Science (COMETS) databases and the associated COMETSbeta and COMETSandSTARS databases © Lynne G. Zucker and Michael R. Darby. All rights reserved. This paper is a part of the NBER's research program in Productivity, Innovation and Entrepreneurship. Any opinions expressed are those of the authors and not those of their employers or the National Bureau of Economic Research.

Communitywide Database Designs for Tracking Innovation Impact: COMETS, STARS and Nanobank

Lynne G. Zucker, Michael R. Darby and Jason Fong
NBER Working Paper No. 17404
September 2011 – Revised March 2014.
JEL No. *O31, O33, O34, O38, C81, J44, J61, J62, M13*

ABSTRACT

Data availability is arguably the greatest impediment to advancing the science of science and innovation policy and practice (SciSIPP). This paper describes the contents, methodology and use of the public online COMETS (Connecting Outcome Measures in Entrepreneurship Technology and Science) database spanning all sciences, technologies, and high-tech industries; its parent COMETSandSTARS database which adds more data at organization and individual scientist-inventor-entrepreneur level restricted by vendor licenses to onsite use at NBER and/or UCLA; and their prototype Nanobank covering only nano-scale sciences and technologies. Some or all of these databases include or will include: US patents (granted and applications); NIH, NSF, SBIR, STTR Grants; Thomson Reuters Web of Knowledge; ISI Highly Cited; US doctoral dissertations; IPEDS/HEGIS universities; all firms and other organizations which ever publish in ISI listed journals beginning in 1981, are assigned US patents (from 1975), or are listed on a covered grant; additional nanotechnology firms based on web search. Ticker/CUSIP codes enable linking public firms to the major databases covering them. A major matching/disambiguation effort assigns unique identifiers for an organization or individual so that their appearances are linked within and across the constituent legacy databases. Extensive geographic coding enables analysis at country, region, state, county, or city levels. The databases provide very flexible sources of data for serious research on many issues in the study of organizations in innovation systems in the development and spread of knowledge, and the economics of science. Enabling the study of these topics, among others, COMETS contributes substantially to the science of science and technology.

Lynne G. Zucker
Professor of Sociology & Public Policy
Director, Center for International Science,
Technology, and Cultural Policy, LSPA
Department of Sociology
University of California, Los Angeles
Los Angeles, CA 90095-1551
and NBER
zucker@ucla.edu

Michael R. Darby
W.E. Cordner Distinguished Professor of Money
& Financial Markets in the UCLA Departments of
Management, Economics, & Public Policy
Anderson Graduate School of Management
University of California, Los Angeles
Los Angeles, CA 90095-1481
and NBER
darby@ucla.edu

Jason Fong,
Assistant Director, Center for International
Science, Technology, and Cultural Policy
UCLA Luskin School of Public Affairs
University of California, Los Angeles
Los Angeles, CA 90095-1656
Jfong@ucla.edu

Communitywide Database Designs for Tracking Innovation Impact: COMETS, STARS and Nanobank*

Lynne G. Zucker, Michael R. Darby and Jason Fong

1 Introduction

For two decades Zucker and Darby, their team and co-authors, as well as other authors and teams working along related lines have been developing a methodology, technology, and the underlying databases required to trace the creation and transmission of new scientifically and/or commercially valuable knowledge, processes, and technologies at the level of country or region; firm (university, government laboratory, or other organization); or individual scientist or engineer (hereafter “scientist” is used to encompass engineers). Data availability is arguably the greatest impediment to advancing the science of science and technology (Zucker and Darby 2011). Since 2003, the Zucker and Darby team has been engaged in a major effort to create increasingly large-scale and comprehensive databases for use of the S&T research community, with intention to enable much wider use of detailed micro-data capable of distinguishing among important competing hypotheses.

This work has evolved into 4 distinct but related databases: Nanobank; COMETS; COMETSbeta; and COMETSandSTARS as summarized in Figure 1. Nanobank contains only records which we have identified as related to nano-scale science and technology (detailed below).

* The construction of Nanobank was supported under major grants from the National Science Foundation (SES-0304727 and SES-0531146) and the University of California’s Industry-University Cooperative Research Program (PP9902, P00-04, P01-02, and P03-01). Additional support was received from the California NanoSystems Institute, Sun Microsystems, Inc., UCLA’s International Institute, and from the UCLA Anderson School’s Center for International Business Education and Research (CIBER) and the Harold Price Center for Entrepreneurial Studies. The COMETS database (also known as the Science and Technology Agents of Revolution or STARS database) is being constructed for public research use under major grants from the Ewing Marion Kauffman Foundation (2008-0028 and 2008-0031) and the Science of Science and Innovation Policy (SciSIP) Program at the National Science Foundation (grants SES-0830983 and SES-1158907) with support from other agencies. Our colleague Jonathan Furner of the UCLA Department of Information Studies played a leading role in developing the methodology for selecting records for Nanobank. We are indebted to our scientific and policy advisors Roy Doumani, James R. Heath, Evelyn Hu, Carlo Montemagno, Roger Noll, and Fraser Stoddart, and to our research team, especially Amarita Natt, Hsing-Hau Chen, Robert Liu, Hongyan Ma, Emre Uyar, and Stephanie Hwang Der. Additional support for COMETS has been provided by The Ewing Marion Kauffman Foundation which both hosts the main COMETS site and in the past ran a COMETS Travel Grants Program to support early use of the COMETS data and presentation of resulting empirical research at conferences through direct grants to users. Certain data included herein are derived from the High Impact Papers, Science Citation Index Expanded, U.S. State Indicators, and U.S. University Indicators of the Institute for Scientific Information®, Inc. (ISI®), Philadelphia, Pennsylvania, USA: © Copyright Institute for Scientific Information®, Inc. 2000-2003. All rights reserved. Certain data included herein are derived from the Nanobank and the Connecting Outcome Measures of Entrepreneurship, Technology, and Science (COMETS) databases and the associated COMETSbeta and COMETSandSTARS databases © Lynne G. Zucker and Michael R. Darby. All rights reserved. This paper is a part of the NBER's research program in Productivity, Innovation and Entrepreneurship. Any opinions expressed are those of the authors and not those of their employers or the National Bureau of Economic Research.

Figure 1. Contents of Nanobank, COMETS, COMETSbeta and, COMETSandSTARS

Note: Items in red are planned future improvements subject to funding and/or vendor's approvals.

Public Library
[no restrictions for
COMETS; Nanobank
for noncommercial use
only]

COMETS
link

Covers:	All science & technology All high-tech firms
Features:	Organization & person matching (same IDs in all databases); sophisticated geography coding including lat. & long. for use in distance function; many start-up/private firms; org types coded
Spans:	US Patents, universities (HEGIS/IPEDS), NIH, NSF, SBIR, STTR Grants, ISI Highly Cited (www.isihighlycited.com) All firms & other orgs. which ever are assigned US patent, or listed on a covered grant
Levels of Data:	US Regions (BEA), Countries; Organizations; Highly Cited Scientists
Years:	Up to 1975-2012

Nanobank
www.nanobank.org

Covers:	Nano-scale science & technology Nanotechnology firms
Features:	Organization & person matching (same IDs in all databases); sophisticated geography coding including lat. & long. for use in distance function; many start-up/private firms; org types coded
Spans:	US Patents; NIH, NSF, SBIR, STTR Grants; Research articles authors, addresses, titles, & sources from Thomson Reuters Web of Knowledge; ISI Highly Cited (www.isihighlycited.com); All nanotech firms & other orgs. which ever publish in ISI listed journals up to 2005, are assigned US patent, or listed on a covered grant; Additional nanotech firms based on web search
Levels of Data:	US Regions (BEA), Countries; Organizations; Highly Cited Scientists
Years:	Up to 1975-2012

Beta-Test Site for
Additions & Updates
[no restrictions for
COMETS; Nanobank
for noncommercial use
only]

COMETS and STARS
needs to migrate to: CometsStars.net

Covers:	All science & technology All high-tech firms New data prior to release + COMETS & Nanobank
Features:	All features of COMETS & Nanobank except nano articles + early access for sophisticated users to new data elements; additional private firms
Spans:	US Patents, universities (HEGIS/IPEDS), NIH, NSF, SBIR, STTR Grants ISI Highly Cited (www.isihighlycited.com) All firms & other org which ever publish in ISI listed journal, are assigned US patent, or listed on a covered grant; additional nanotech firms based on web search
Levels of Data:	US Regions (BEA), Countries; Organizations; Highly Cited Scientists
Years:	Up to 1975-2012

Confidential
Files for On-Site
Use Only
[NBER Productivity
Group Members and
Visiting Scholars - both
by Individual Application]

COMETS-NBER

Covers:	All science & technology All high-tech firms
Features:	All features of COMETS + Analysis datasets at levels of region and country, organization (eg, firm), and individual scientists
Spans:	US Patents, universities (HEGIS/IPEDS), NIH, NSF, SBIR, STTR Grants, ISI Highly Cited (www.isihighlycited.com) All firms & other orgs. which ever are assigned US patent, or listed on a covered grant
Levels of Data:	US Regions (BEA), Countries; Organizations; Highly Cited Scientists
Years:	Up to 1975-2012

Nanobank-NBER

Covers:	Nano-scale science & technology Nanotechnology firms
Features:	All features of Nanobank + Analysis datasets at levels of region and country, organization (eg, firm), and individual scientists
Spans:	US Patents, NIH, NSF, SBIR, STTR Grants Thomson Reuters Web of Knowledge ISI Highly Cited (www.isihighlycited.com) All nanotech firms & other org which ever publish in ISI listed journals up to 2005, are assigned US patent, or listed on a covered grant; additional nanotech firms based on web search
Levels of Data:	US Regions (BEA), Countries; Organizations; Highly Cited Scientists
Years:	Up to 1975-2012

Confidential
Files for On-Site
Use Only
[Individual Applicants
as Fellows of the
Center for International
Science, Technology
and Cultural Polcy,
UCLA Luskin School
of Public Affairs]

COMETS-UCLA

Covers:	All science & technology All high-tech firms
Features:	All features of COMETS-NBER + Web of Knowledge articles database with person matching IDs for individual scientists
Spans:	US Patents, universities (HEGIS/IPEDS), NIH, NSF, SBIR, STTR Grants, ISI Highly Cited (www.isihighlycited.com) All firms & other org which ever publish in ISI listed journal, are assigned US patent, or listed on a covered grant
Levels of Data:	US Regions (BEA), Countries; Organizations; Highly Cited Scientists
Years:	Up to 1975-2012

Nanobank-UCLA

Covers:	Nano-scale science & technology Nanotechnology firms
Features:	All features of Nanobank-NBER + Web of Knowledge articles database with person matching IDs for individual scientists
Spans:	US Patents, NIH, NSF, SBIR, STTR Grants Thomson Reuters Web of Knowledge ISI Highly Cited (www.isihighlycited.com) All nanotech firms & other org which ever publish in ISI listed journals up to 2005, are assigned US patent, or listed on a covered grant; additional nanotech firms based on web search
Levels of Data:	US Regions (BEA), Countries; Organizations; Highly Cited Scientists
Years:	Up to 1975-2012

Early, beta-test releases of Nanobank provided an important source of data for some of the other articles in this issue. Nanobank also served as a prototype and test-bed for the Science and Technology Agents of Revolution (STARS) project which extended coverage to all areas of science and engineering and all high-tech industries as well as extending both the period of coverage (from an end date in Nanobank of 2004 to an end date of 2010 except for article data, due to licensing restrictions which limit it to 2004) and the heritage databases included (e.g., adding NIH, NSF, SBIR. and STTR grants). Although the coverage extended far beyond the top scientists, the STARS name linked in to Zucker and Darby's pioneering work on the role of star scientists in high-tech firm formation and success, focused primarily on biotechnology and nanotechnology (Zucker and Darby 1996, 2009; Zucker, Darby and Brewer 1998; Zucker, Darby and Armstrong 1998; Darby and Zucker 2005, 2007).[1] This difference and potential confusion with the recent Federal STARmetrics program, led us to name the new public database COMETS (Connecting Outcome Measures in Entrepreneurship Technology and Science). COMETS is hosted by the Ewing Marion Kauffman Foundation at www1.kauffman.org/comets. This paper discusses the major elements included and challenges overcome in construction of each database in turn, beginning with Nanobank.

2 Nanobank

The Nanobank database is available at www.nanobank.org for free use for research purposes. You will be asked to send a brief statement of your planned use to zucker@ucla.edu. Covering nano-scale science and engineering as well as its commercialization, Nanobank serves as an enabling or platform technology for social science, business, and policy research on the science origins of nanotechnology and its commercialization. Of special relevance is a system of unique ID numbers for firms, universities and other organizations used as they appear within and across such components as journal articles, US patents, and NSF and NIH grants. Nanobank will be archived by the Zucker-Darby team at the Center for International Science, Technology and Cultural Policy (CISTCP) in the UCLA Luskin School of Public Affairs and at the National Bureau of Economic Research. It will be extended and updated as an integrated component of the COMETS database described in Section 3 below.[2] A similar system of unique ID numbers for frequently publishing and/or patenting individuals will be completed as part of the STAR/COMETS database project and included in COMETSandSTARS in the future. Table 1 provides an overview of the data currently available in Nanobank. Perusing the list of parsed fields gives an idea of the rich variables included and which can be constructed using Nanobank.

[1] Liebeskind, Oliver, Zucker and Brewer (1996), Zucker and Darby (1997), and Zucker, Darby and Armstrong (2002) all offer evidence that while star-scientist employees and collaborators have the biggest impacts, other scientists also make important contribution whether as employees of the firm or networked to them.

[2] The COMETS database covers all science and technology fields and all high-technology areas. Nanotechnology flags are included for those documents identified as such using the Nanobank methodologies described below. This identification is extended as permitted by funding. Nanobank *per se* cannot be extended due to licensing restrictions, limiting the article data to through 2004.

Three principal technical challenges were overcome in constructing these databases: (a) defining nano-scale science and engineering and its commercial applications; (b) matching appearances of the same organization within and across the component databases; and (c) locating

Table 1. Nanobank Data Description from Nanobank.org as of August 11, 2011[a]

Table Name	Column Name	Col. Type	Column Description	Reference Sheet
SECTION 1 : Articles[a]				
articles	article_id	integer	article ID	
	journal_id	integer	journal ID	
	article_title	character	article title	
	journal_title	character	journal title	
	volume	character	volume number	
	issue	character	issue number	
	bpage	character	beginning page	
	epage	character	ending page	
	pub_year	integer	publication year	
	pub_date	character	publication date	
	authority_flag	boolean	1 if article is in the authority set, 0 otherwise	NANO Identification
	nanobank_flag	boolean	1 if article is in the Nanobank identification set, 0 otherwise	NANO Identification
article_authors	article_id	integer	article ID	
	pos	integer	order of appearance of this author	
	lastname	character	last name	
	first_init	character	first initial	
	middle_inits	character	middle initials (possibly multiple initials)	
article_reprint_addrs	article_reprint_addr_id	integer	reprint address ID	
	addr_author	character	corresponding author name	
	org_name	character	name of university, company, institution, etc.	
	org_subname	character	name of suborganization, department, etc.	
	org_id	character	alphanumeric code specific to each organization	Org Codes Info
	org_type	character	organization type	Org Codes Info
	org_nano_name	character	The non-abbreviated version of the organization name that appears most among the organization's nano-related articles and patents.	Org Codes Info
	full_addr	character	full address	
	street	character	street address	
	city	character	city	
	city_std	character	FIPS standardized city name (for USA)	FIPS 55 Info
	city_code	integer	FIPS city code (for USA)	FIPS 55 Info
	county	character	county	FIPS 55 Info
	county_code	integer	FIPS county code (for USA)	FIPS 55 Info
	state	character	state	
	state_code	integer	FIPS state code (for USA)	FIPS 55 Info
	postal_code	character	postal code	
	country	character	country	
	country_code	character	ISO country code	ISO Country Info
	bea_code	integer	BEA code	BEA Info
	article_id	integer	article ID	
	journal_id	integer	journal ID	
	article_title	character	article title	
	journal_title	character	journal title	
	bpage	character	beginning page	
	epage	character	ending page	
	volume	character	volume number	
	issue	character	issue number	
	pub_year	integer	publication year	
	pub_date	character	publication date	

Table 1. Nanobank Data Description from Nanobank.org as of August 11, 2011 (continued)

Table Name	Column Name	Col. Type	Column Description	Reference Sheet
article_other_addrs	article_other_addr_id	integer	other (non-reprint) address ID	
	org_name	character	name of university, company, institution, etc.	
	org_subname	character	name of suborganization, department, etc.	
	org_id	character	alphanumeric code specific to each organization	Org Codes Info
	org_type	character	organization type	Org Codes Info
	org_nano_name	character	The non-abbreviated version of the organization name that appears most among the organization's nano-related articles and patents.	Org Codes Info
	full_addr	character	full address	
	street	character	street address	
	city	character	city	
	city_code	integer	FIPS city code (for USA)	FIPS 55 Info
	county	character	county	FIPS 55 Info
	county_code	integer	FIPS county code (for USA)	FIPS 55 Info
	state	character	state	
	state_code	integer	FIPS state code (for USA)	FIPS 55 Info
	postal_code	character	postal code	
	country	character	country	
	country_code	integer	ISO country code	ISO Country Info
	bea_code	integer	BEA code	BEA Info
	article_id	integer	article ID	
	journal_id	integer	journal ID	
	article_title	character	article title	
	journal_title	character	journal title	
	bpage	character	beginning page	
	epage	character	ending page	
	volume	character	volume number	
	issue	character	issue number	
	pub_year	integer	publication year	
	pub_date	character	publication date	
SECTION 2 : Patents				
patents	patent_id	integer	patent number	
	num_claims	integer	number of claims	
	grant_date	date	grant date	
	app_num	character	application number	
	app_date	date	application date	
	patent_title	character	patent title	
	authority_flag	boolean	1 if patent is in the authority set, 0 otherwise	NANO Identification
	nanobank_flag	boolean	1 if patent is in the Nanobank identification set, 0 otherwise	NANO Identification
patent_citations	patent_id	integer	patent number	
	year	integer	grant year	
	citations	integer	# of patents granted this year that cite this patent	
patent_int_classes	patent_id	integer	patent number	
	pos	integer	order of appearance for this class	
	intl_class	character	international patent class	
patent_US_classes	patent_id	integer	patent number	
	pos	integer	order of appearance for this class	
	us_class	character	US patent class	
patent_abstracts	patent_id	integer	patent number	
	patent_title	character	patent title	
	patent_abstract	character	patent abstract	

Table 1. Nanobank Data Description from Nanobank.org as of August 11, 2011 (continued)

Table Name	Column Name	Col. Type	Column Description	Reference Sheet
patent_assignees	patent_id	integer	patent number	
	pos	integer	order of appearance for this assignee	
	org_name	character	name of university, company, institution, etc.	
	org_id	character	alphanumeric code specific to each organization	Org Codes Info
	org_type	character	organization type	Org Codes Info
	org_nano_name	character	The non-abbreviated version of the organization name that appears most among the organization's nano-related articles and patents.	Org Codes Info
	city	character	city	
	city_std	character	FIPS standardized city name (for USA)	FIPS 55 Info
	city_code	integer	FIPS city code (for USA)	FIPS 55 Info
	county	character	county	FIPS 55 Info
	county_code	integer	FIPS county code (for USA)	FIPS 55 Info
	state	character	state	
	state_code	integer	FIPS state code (for USA)	FIPS 55 Info
	country	character	country	
	country_code	integer	ISO country code	ISO Country Info
	bea_code	integer	BEA code	BEA Info
	num_claims	integer	number of claims	
	granted	date	grant date	
	appnum	character	application number	
	applied	date	application date	
	patent_title	character	patent title	
patent_inventors	patent_id	integer	patent number	
	pos	integer	order of appearance for this inventor	
	last_name	character	last name	
	first_name	character	first name	
	middle_name	character	middle name	
	suffix	character	name suffix	
	street	character	street address	
	city	character	city	
	city_std	character	FIPS standardized city name (for USA)	FIPS 55 Info
	city_code	integer	FIPS city code (for USA)	FIPS 55 Info
	county	character	county	
	county_code	integer	FIPS county code (for USA)	FIPS 55 Info
	state	character	state	
	state_code	integer	FIPS state code (for USA)	FIPS 55 Info
	postal_code	character	postal code	
	country	character	country	
	country_code	integer	ISO country code	ISO Country Info
	bea_code	integer	BEA code	BEA Info
	num_claims	integer	number of claims	
	granted	date	grant date	
	appnum	character	application number	
	applied	date	application date	
	patent_title	character	patent title	
SECTION 3 : Grants[b]				
grants	grant_id	integer	grant ID	
	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	fiscal_year	integer	fiscal year	
	start_date	date	start date	
	end_date	date	end date	
	last_amend_date	date	last amendment date	
	instrument	character	award instrument	
	amount	integer	award amount	
	grant_title	character	grant title	
	authority_flag	boolean	1 if grant is in the authority set, 0 otherwise	NANO Identification
	nanobank_flag	boolean	1 if grant is in the Nanobank identification set, 0 otherwise	NANO Identification

Table 1. Nanobank Data Description from Nanobank.org as of August 11, 2011 (concluded)

Table Name	Column Name	Col. Type	Column Description	Reference Sheet
grantee_orgs	grant_id	integer	grant ID	
	grant_agency	character	granting agency	
	org_name	character	name of university, company, institution, etc.	
	org_id	character	alphanumeric code specific to each organization	Org Codes Info
	org_type	character	organization type	Org Codes Info
	org_nano_name	character	The non-abbreviated version of the organization name that appears most among the organization's nano-related articles and patents.	Org Codes Info
	street	character	street address	
	city	character	city	
	city_std	character	FIPS standardized city name (for USA)	FIPS 55 Info
	city_code	integer	FIPS city code (for USA)	FIPS 55 Info
	county	character	county	
	county_code	integer	FIPS county code (for USA)	FIPS 55 Info
	state	character	state	
	state_code	integer	FIPS state code (for USA)	FIPS 55 Info
	postal_code	character	postal code	
	country	character	country	
	country_code	character	ISO country code	ISO Country Info
	bea_code	integer	BEA code	BEA Info
grant_pis	grant_id	integer	grant ID	
	grant_agency	character	granting agency	
	last_name	character	PI last name	
	first_name	character	PI first name	
	middle_name	character	PI middle name	
grant_abstracts	grant_id	integer	grant ID	
	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	grant_title	character	grant title	
	grant_abstract	character	grant abstract	
SECTION 3.1 : NSF-specific Grant Information				
grant_nsf	grant_id	character	grant ID	
	prog_manager	character	program manager	
	directorate	character	NSF directorate	
grant_nsf_programs	grant_id	integer	grant ID	
	pos	integer	order of appearance of this program	
	program	character	NSF program name	
grant_nsf_fields	grant_id	integer	grant ID	
	pos	integer	order of appearance of this field of application	
	field	character	NSF field of application	
	field_code	character	code for this field of application	
grant_nsf_co_pis	grant_id	integer	grant ID	
	pos	integer	order of appearance of this co-PI	
	lastname	character	co-PI last name	
	firstname	character	co-PI first name	
	middlename	character	co-PI middle name	
SECTION 3.2 : NIH-specific Grant Information				
grant_nih	grant_id	integer	grant ID	
	nih_icd	character	NIH institute, center, or division	
	nih_irg	character	NIH initial review group	
grant_nih_tags	grant_id	integer	grant ID	
	pos	integer	order of appearance of this tag	
	tag	character	NIH descriptive tag	

Source: Extract from full file downloadable at “Nanobank codebook” at http://www.nanobank.org/.
Notes: Reference sheet refers user to sources and detailed coding on another Excel worksheet.

[a] Article data are permanently frozen at 2004 given our inability to license additional data for public deployment after our initial agreement with the Institute for Scientific Information (ISI) that covered only nanoscience articles identified by us up to 2004.

[b] Grants added in 2011. We have updated grants to improve the accuracy of the merger of two legacy datasets provided by NIH on February 12, 1914.

geographically addresses listed on the documents. The nature of these challenges and how we solved them are the subjects of Sections 2.1, 2.2, and 2.3 below, respectively.

2.1. Defining Nanotechnology Operationally

Nanobank is a digital library containing a collection of documents related to various topics in the nanotechnology field. These documents include scientific journal articles, patents, and government grants. This database contains bibliographic information, including titles, abstracts, publication years, and author names. Information on associated organizations is also provided. This includes unique IDs for each distinct organization and geocoding information for their locations as discussed in Sections I.B and I.C, respectively.

2.1.1 Data Sources and Content

The journal articles portion of Nanobank contains 580,711 articles from peer reviewed journals. The sources of this data are the Science Citation Index, Arts & Humanities Citation Index, and Social Sciences Citation Index of the Institute for Scientific Information Inc. (ISI). These sources contain a total of over 25,000,000 articles from over 8,700 peer reviewed scientific journals. Nanobank contains the subset of articles that are determined as described below to be relevant to nanotechnology. The article data includes unique ID numbers for each article, article titles, abstracts, journal volume number, journal issue number, publication year, author names, and the names and addresses of organizations affiliated with the authors.[3]

The patent data in Nanobank includes 240,437 patents filed with the United States Patent and Trademark Office (USPTO). The source of this data is a number of flat text files which are made available by the USPTO. The files used for Nanobank contain data on over 4,000,000 patents with grant years ranging from 1976 to 2005. Nanobank contains the subset of patents that are determined as described below to be relevant to nanotechnology. The patent data includes USPTO patent ID numbers, patent titles, abstracts, U.S. and international patent classifications, application dates, grant dates, and the names and addresses of inventors and assignees.

The government grants data includes 52,830 grants, with 29,541 coming from the National Institutes of Health (NIH) and 23,289 from the National Science Foundation (NSF) for 1972-2006. This data includes the ID numbers assigned by the grant agency, titles, abstracts, PI names, co-PI names, grant amounts, and receiving organization names and addresses.

2.1.2 Document Selection

As is normal with an emerging field with contested boundaries, there is no clear definition of nanotechnology in any of the legacy databases integrated in Nanobank. The documents selected for Nanobank represent our best efforts to include – to the extent possible given automated search – all documents which might be viewed by a significant number of experts as relevant to nano-scale science and engineering and its commercial applications. We made the conscious decision to err on

[3] What information can be made public through Nanobank is limited by the terms of our license from ISI (now merged into Thomson Reuters and deployed on the Web of Knowledge) and other vendors of component proprietary databases. Our article data is limited to 2004. We cannot include the ISI ID code for an article or counts or links of citations to it, but we do format the journal citation to exactly match those used by ISI so that those with access to ISI data can link readily to that database.

the side of inclusion, and some users will choose to select subsets more attuned to their particular operational definitions.

Three methods are used to determine which documents are nanotechnology-relevant. The first “keywords” or Boolean method is based on the existence of one or more specified words or phrases found in the available text portions of the documents. Titles, abstracts, and keywords were available for articles and grants. The full text was available for patents. The keywords method searches for text patterns which match words or phrases related to nanotechnology. A drawback to this method is that it is less effective for very early or very recent documents. This is the case because early documents were written before the search terms were in common use, and recent documents have terms that are too new to be included in the search terms. Our keywords were any term that was prefixed with “nano” and (A) the 140 most commonly occurring noun phrases in the *Virtual Journal of Nanoscale Science & Technology* (*VJN*), (B) 297 “glossary” terms primarily derived from recommended search lists received from collaborators and advisory board members who are specialists in the field and supplemented by a web search of nanotechnology glossaries, (C) with the exception of pure measurement terms. The 140 most commonly occurring noun phrases in VJN articles up through 2003 was found by using a tool called the Apple Pie Phraser (APP) which is a tool that analyzes the grammatical structure of a sentence and identifies the noun phrase(s) in the sentence. Table 2 lists the keywords (other than “nano*”) used in constructing Nanobank in the form of regular expression text patterns, so a single entry could represent a number of possible terms. The terms in Part C of the table are the pure measurement terms which were excluded from triggering selection of a record as nanotechnology relevant.

The second “probabilistic” document selection method is a relative frequency method which selects some of the same and some additional documents to complement the Boolean method and fill in for some of its shortcomings is due to Jonathan Furner and Hongyan Ma, as implemented by Jason Fong in selection of cut points in the probability distribution for nano-identification. This probabilistic method analyzes the document text and ranks the documents in order of relevance to a set of query terms. However, unlike the keywords method, this set of query terms is not preselected. The query terms used for the probabilistic method adapt to the contents of the document set. This allows the inclusion of terms that have not been previously identified as nanotechnology- relevant.

Since the search terms are not preselected for the probabilistic method, a process is needed for automatically generating a set of nanotechnology- relevant terms. The Xapian search engine library is used for performing the ranking calculations needed for the term selection process. First, an initial set of query terms is derived from the text of the articles in the *Virtual Journal of Nanoscale Science & Technology* (VJN). This initial set is created by assuming that all of the documents in VJN are relevant, and then selecting the terms that Xapian determines to be the highest ranked for the purpose of characterizing the VJN documents – i.e., those that are relatively common in VJN articles relative to their frequency in the universe of all articles. This set of search terms is used to select an initial set of relevant documents from the full data set. Additional highly ranked terms are then chosen from this initial set of relevant documents. These additional terms are added to the search terms and an expanded set of relevant documents are selected from the full data set. This expanded set of relevant documents is used for Nanobank.

Table 2. Keywords Used in Nanobank Document Selection

Part A - Terms Based on the *Virtual Journal of Nanoscale Science & Technology*

(c|carbon).nanotube.field.emitter
dip.pen.nanolithography
gaas.quantum.dot.laser
mesoscopic.structur\w*
nanoring
ni.nanowire
quantum.hall.effect
(gan|gallium.nitride).nanowire
molecular.nanomagnet
nanotube.axi
quantum.dynamic
quantum.interference.effect
semiconduct\w*.nanowire
semiconduct\w*.quantum.wire
aln.quantum.dot
mesoscopic.superconduct\w*
nanodot
single.?quantum.well
(gan|gallium.nitride).quantum.dot
mesoscopic.ring
multiple.?quantum.well
quantum.teleportation
quantum.well.structur\w*
semiconduct\w*.nanocrystal\w*
zigzag.nanotube
cds.nanocrystal\w*
ge.nanocrystal\w*
mesoscopic
nanotube.bundle
nanotube.diameter
quantum.dot.structur\w*
quantum.gate
quantum.tunneling
(c|carbon).nanotube.bundle
chaotic.quantum.dot
mesoscopic.fluctuation
nanodevice
nanoelectromechanic\w*.system
quantum.dot.superlattice
(c|carbon).nanoparticle
(c|carbon).nanostructur\w*
algaas.quantum.well
magnetic.nanoparticle
open.quantum.dot
quantum.bit
two.quantum.dot
(au|gold).nanowire

nanofabrication
nanomaterial
quantum.efficiency
quantum.fluctuation
quantum.coherence
quantum.hall.regime
quantum.information
quantum.conduct\w*
quantum.dot.system
(si|silicon).nanostructur\w*
double.?wall\w*.(c|carbon).nanotube
quantum.effect
quantum.communication
semiconduct\w*.(c|carbon).nanotube
superconduct\w*.quantum.interference.device
zno.nanowire
magnetic.nanostructur\w*
metal\w*.nanowire
nanotechnology
cdse.nanocrystal\w*
nanocrystal\w*.(si|silicon)
nanobelt
quantum.well.state
semiconduct\w*.nanotube
spherical.quantum.dot
metal\w*.(c|carbon).nanotube
nanomagnet
nanocomposite
nanodiamond
nanorod
quantum.confin\w*.effect
quantum.entanglement
quantum.dot.array
(ag|silver).nanoparticle
metal\w*.nanotube
single.?wall\w*.(c|carbon).nanotube.bundle
quantum.phase.transition
quantum.ring
quantum.system
quantum.transport
semiconduct\w*.quantum.well
nanoimprint.lithography
zeolite
inp.quantum.dot
nanofiber
quantum.mechanic\w*
(c|carbon).nanofiber

quantum.dot.laser
semiconduct\w*.nanostructur\w*
multi.?wall\w*.nanotube
nanocontact
quantum.interference
cdse.quantum.dot
ingaas.quantum.dot
quantum.cascade.laser
quantum.information.process\w*
(si|silicon).quantum.dot
single.?wall\w*.nanotube
spintronic
ge.quantum.dot
coupled.quantum.dot
nanopore
mesoscopic.system
metal\w*.nanoparticle
(si|silicon).nanowire
(au|gold).nanoparticle
double.?quantum.dot
quantum.point.contact
(bn|boron.nitride).nanotube
quantum.state
nanocluster
wall\w*.(c|carbon).nanotube
nanoscale
single.?quantum.dot
gaas.quantum.well
quantum.confin\w*
nanoindent\w*
semiconduct\w*.quantum.dot
(si|silicon).nanocrystal\w*
quantum.comput\w*
gaas.quantum.dot
quantum.wire
inas.quantum.dot
nanostructur\w*
nanocrystal\w*
multi.?wall\w*.(c|carbon).nanotube
nanowire
nanoparticle
quantum.well
single.?wall\w*.(c|carbon).nanotube
nanotube
(c|carbon).nanotube
quantum.dot

Part B - Terms Derived from Nanotechnology Glossaries

asia.pacific.nanotechnology.forum
atomic.force.microscop\w*
atomic.manipulation
atomic.resolution
auger.electron
auger.electron.spectroscopy
bio.assembl
biofabrication
biomedical.nanotechnology
biomimetic
biomimetic.chemistry
biomimetic.material
biomimetic.synthesis

molecular.motor
molecular.nanogenerator
molecular.nanoscience
molecular.nanotechnology
molecular.repair
molecular.robotic
molecular.scale.manufacturing
molecular.sieve
molecular.surgery
molecular.switch
molecular.systems.engineering
molecular.technology
molecular.wire

nanorod
nanorope
nanoscale
nanoscale.self.assembly
nanoscale.synthesis
nanoscience
nanoscopic.scale
nanosensor
nanoshell
nanosource
nanosphere
nanostructure
nanostructured.surface

Table 2. Keywords Used in Nanobank Document Selection (continued)

Part B - Terms Derived from Nanotechnology Glossaries (continued)

biomolecular.assembl
biomolecular.nanoscale.computing
biomolecular.nanotechnology
bionanotechnology
bionems
blue.goo
bottom.up.nanotechnology
brownian.assembly
buckminsterfullerene
bucky.ball
buckyball
buckytube
c60
c60.molecule
cantilever.tip
carbon.nanofoam
carbon.nanotube
cascade.molecule
cell.pharmacology
cell.repair.machine
cell.surgery
cognotechnology
computational.nanotechnology
computronium
conductance.quantization
convergent.assembly
cryogenic.afm
dendrimer
dig.pen.nanolithography
dip.pen.nanolithography
directed.assembler
disassembler
dna.chip
dna.computing
dopeyball
dry.nanotechnology
electron.beam.lithography
electron.transport.chain
electrostatic.force.microscop\w*
epitaxial.film
epitaxy
european.nanobusiness.association
fat.fingers.problem
femtoengineering
femtotechnology
fluidic.self.assembly
fullerene
giant.magnetoresistance
glycodendrimer
glyconanotechnology
gnr.technologies
golden.goo
gray.goo
green.goo
grey.goo
gripper
immune.machine
institute.for.molecular.manufacturing
khaki.goo
lab.on.a.chip
langmuir.blodgett
laser.tweezer
lateral.force.microscop\w*
limited.assembler
limited.molecular.nanotechnology
lofstrom.loop
low.dimension.structure
low.dimensional.structure

moletronic
molmac
monomolecular.computing
multiwalled.nanotube
nanarchist
nanarchy
nanite
nano.assembly
nano.cubic.technology
nano.lithography
nano.optic
nano.pollution
nano.warfare
nanoarray
nanoassembler
nanobarcode
nanobarcodes.particle
nanobiology
nanobioprocessor
nanobiotechnology
nanobiotechnology.platform
nanobot
nanobubble
nanobusiness.alliance
nanobuzz
nanocatalysis
nanochemistry
nanochip
nanocircle
nanocluster
nanocomposite
nanocomputer
nanocone
nanocrystal
nanocrystal.antenna
nanodefense
nanodentistry
nanodetector
nanodevice
nanodisaster
nanodot
nanoelectromechanical.system
nanoelectronic
nanoelectrospray
nanoengineering
nanofabrication
nanofactory
nanofacture
nanofiber
nanofibre
nanofiltration
nanofluidic
nanofoam
nanogate
nanogear
nanogenomic
nanogypsy
nanohacking
nanoimaging
nanoimprint.lithography
nanoimprint.machine
nanoimprinting
nanoindentation
nanolabel
nanolithography
nanomachine
nanomanipulation
nanomanipulator

nanosurgery
nanoswarm
nanosystem
nanotechism
nanotechnology
nanoterrorism
nanotube
nanowalker
nanowetting
nanowire
national.nanotechnology.initiative
optical.trapping
optical.tunneling
optical.tweezer
organic.led
peptide.nanotube
phantom
pico.technology
picoengineering
pink.goo
polymorphic.smart.material
positional.assembly
poss.nanotechnology
protein.design
protein.engineering
proximal.probe
quantum.computation
quantum.computer
quantum.computing
quantum.confined.atom
quantum.cryptography
quantum.dot
quantum.dot.nanocrystal
quantum.interferometric.lithography
quantum.mirage
quantum.nanophysic
quantum.well
quantum.wire
quantumbrain
qubit
red.goo
rosette.nanotube
rotaxane
scanning.capacitance.microscop\w*
scanning.electron.microscop\w*
scanning.force.microscop\w*
scanning.near.field.optical.microscop\w*
scanning.probe.lithography
scanning.probe.microscop\w*
scanning.probe.nanolithography
scanning.thermal.microscop\w*
scanning.tunneling.electron.microscop\w*
scanning.tunneling.microscop\w*
self.assembled.monolayer
single.beam.gradient.trap
single.cell.detection
single.cell.manipulation
single.dna.molecule.sequencing
single.electron.device
single.electron.transfer
single.molecule.detection
single.molecule.manipulation
single.walled.carbon.nanotube
smart.material
soft.lithography
spin.coating
spintronic
star.trek.scenario

Table 2. Keywords Used in Nanobank Document Selection (concluded)

Part B - Terms Derived from Nanotechnology Glossaries (concluded)

magnetic.force.microscop\w*
metal.nanoshell
micellar.nanocontainer
microengineering.interfaces.with.living.cell
microfabrication
microfluidic
microfluidic.channel
micromanipulation
microtubule
minatec
molecular.assembler
molecular.beam.epitaxy
molecular.electronic
molecular.integrated.microsystem
molecular.machine
molecular.manipulator
molecular.manufacturing
molecular.mechanic
nanomanufacturing
nanomaterial
nanomechanical
nanomedicine
nanomotor
nanoparticle
nanopgm
nanopharmaceutical
nanophase.carbon.materials
nanophobia
nanophotonic
nanophysic
nanoplumbing
nanopore
nanoporous
nanoprism
nanoprobe
nanoreplicator
stewart.platform
sticky.fingers.problem
substrate
superlattice.nanowire.pattern
technocyte
textronic
thin.film
top.down.nanotechnology
tubeologist
two.dimensional.material
ubergoo
universal.assembler
up.converting.phosphor
utility.fog
vasculoid
virtual.nanomedicine
wet.nanotechnology
zettatechnology

Part C - Excluded Measurement Terms

\bnm\b
angstrom
attomole
femtometer
femtomole
nanogram
nanoliter
nanometer
nanometre
nanonewton
nanosecond
nanovolt
picoliter
picometer
picomole
piconewton
yoctomole
zeptomole
zeptosecond

The third method used for document selection adds documents that are identified as nano-relevant by an outside “authoritative” source. For journal articles, the *Virtual Journal of Nanoscale Science & Technology* is considered to be an authoritative source. Any article found in VJN is also included in the Nanobank dataset. The US Patent Classification 977 is used as an authoritative source for the patent data. This is the classification for nanotechnology assigned by the USPTO. For NIH grants, additional grants were selected when the program name of the grant included “nano.” For NSF grants, additional grants were selected when the descriptive tag of the grant included “nano.”

Table 3. Breakdown of Documents in Nanobank by Selection Criteria as of October 11, 2011

Documents Selection Criteria			Number in Nanobank of			
Keywords	**Probabilistic**	**Authoritative**	**Articles**	**Patents**	**NSF Grants**	**NIH Grants**
Yes	No	No	74876	24669	2668	7621
Yes	No	Yes	1040	232	278	652
Yes	Yes	No	159171	30881	5470	1871
Yes	Yes	Yes	11527	2793	2030	1085
No	Yes	No	328992	180654	11562	17621
No	Yes	Yes	2582	556	232	120
No	No	Yes	2522	651	1049	571
			580710	**240436**	**23289**	**29541**

The documents in Nanobank were selected on the basis of meeting one or more of the three criteria. Table 3 tabulates the number of documents according to type of document and which of the criteria were met for a given document. Clearly the probabilistic method added the most documents with considerably fewer being selected by the keywords method and only a relatively small number identified in any of the authoritative document selections.

Nanobank comprises the union of all documents selected by any one or more of these three methods. The data contain codes permitting users to distinguish between documents that would have been included in the database using either of the first two methods versus those which are included because they are in the specified authoritative sets.

2.2 Matching Organizations within and across Legacy Databases

Each organization found in the Nanobank data is assigned an alphanumeric code. These organization codes are composed of two parts. The first part is a two-character code that identifies the organization's type. Organization types include firms, universities, national labs, research institutes, U.S. government organizations, hospitals, and academies of sciences. The second part of an organization code is a numeric code that uniquely identifies an organization within each type.

The organization codes aid in the grouping of observations of the same organization by standardizing the various forms of an organization name. For example, the name "IBM" can also appear as "IBM Corp." or "IBM Corporation." The word "University" in an organization name can also appear in an abbreviated form as "Univ", or it can appear in another language, for example, as "Universidad." Common misspellings in organization names are also handled by using organization codes as the grouping unit. We made no systematic attempts to capture and trace name changes or to code subparts of organizations which do not incorporate the parent's name in their own.

Combining organization code and address fields can be used to obtain data for organizations at the establishment level – that is, activity of an organization occurring at a particular location. However, such constructed establishment data should be used carefully, since the underlying legacy databases do not use that concept.

Probably the most difficult cases are for US patents, where there is no definitive indication of where or in what organization the inventive activity occurred. Inventors are required by law to be identified by residence address. Organizations appear only if the patent is assigned to them by the inventors by the time the patent is issued ("assignee at issue"). These assignee organizations are most likely to be an employer of one or more of the inventors, but in some cases independent inventors sell the rights to their invention at arm's length to an organization prior to issue of the patent, or indeed inventors' employers can similarly sell their rights to another organization before issuance of the patent. A familiar example of the latter would be the case of a university faculty inventor whose university has given a firm funding the research a right of first refusal to any resulting intellectual property rights. In any case for multi-location organizations, the address of the assignee is often the corporate headquarters and not the location of the inventive activity. Therefore, we recommend using inventor addresses to locate the inventive activity geographically. When the individual ID numbers are available for frequent authors and inventors, it may be possible to infer the extent of error introduced by using empirically the assignee at issue as the employer of the inventors.

2.3 Geocoding of Addresses in Nanobank

A significant amount of geocoding work was performed on the Nanobank dataset to make the geographic information easier to use. The geocoding work had a number of goals, including: standardization between the various naming conventions used in different sources, standardization of non-uniformly recorded data, and correction of common misspellings. For observations with locations in the United States, the geocoding also provides additional grouping units not available in the original source data. For example, city and state information are commonly found in the source data, but our geocoding work adds additional information such as county locations and US Bureau of Economic Analysis (BEA) functional economic areas. Of particular interest is the latitude and longitude associated with each address, permitting easy computation of distances between locations if that variable is of interest.

U.S. observations are those that are located within the 50 U.S. states, the District of Columbia, and 7 U.S. associated areas. Cities, states, and counties are given numeric codes from the "Populated Places" data obtained from the FIPS 55 database. City names are standardized and matched to names in the FIPS database on a state-by-state basis. In the journal articles data, 99.98% of the U.S. observations are assigned a definite city code and state code.

The BEA economic areas are composed of 179 functional economic areas in the U.S. assigned by the Bureau of Economic Analysis. These areas consist of one or more economic nodes – metropolitan or micropolitan statistical areas that serve as regional centers of economic activity and the surrounding counties that are economically related to the nodes. The BEA areas used for Nanobank were defined on November 17, 2004. Each county in the U.S. is assigned to a unique BEA area, with multiple counties contained within each BEA area.

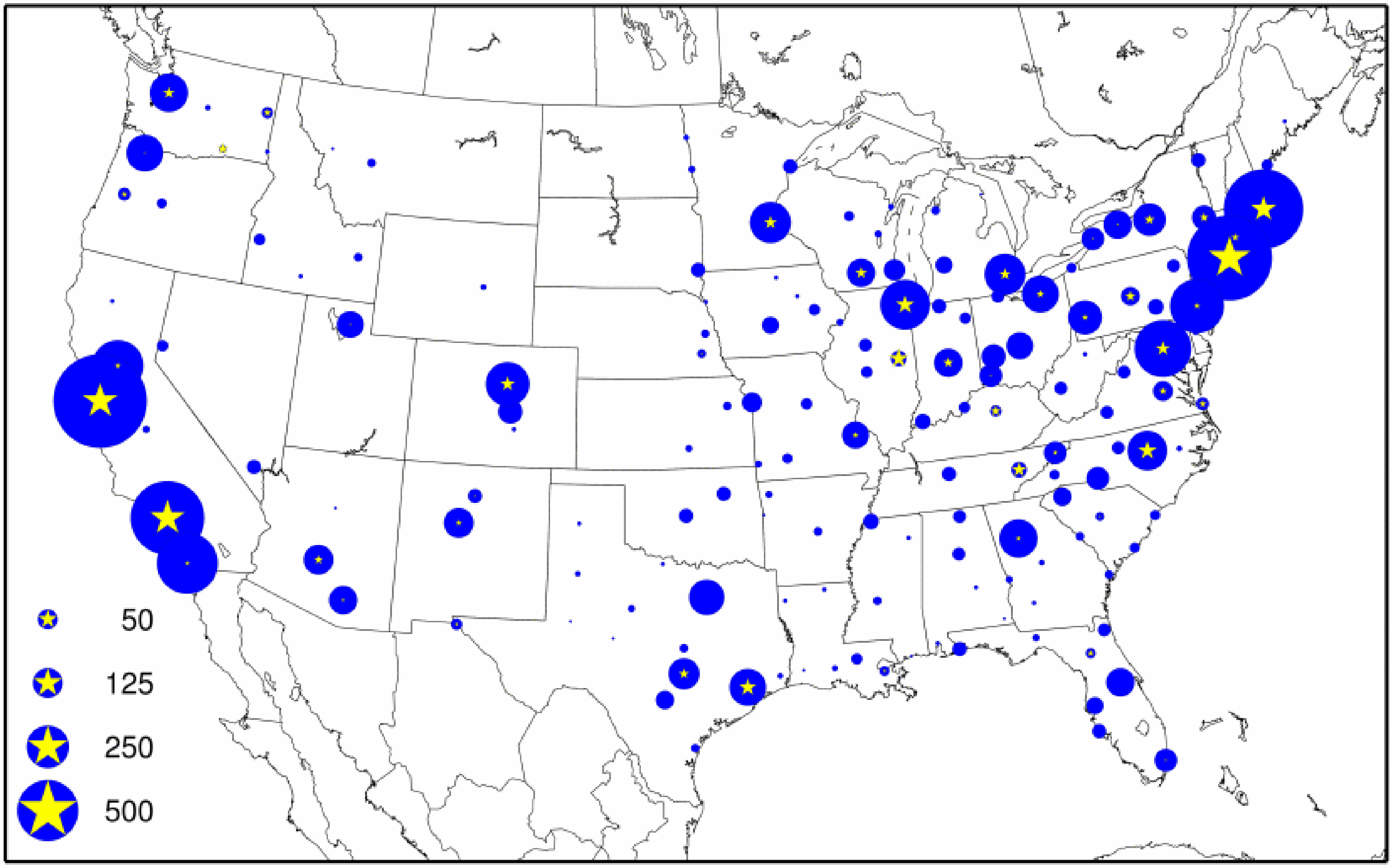
2.4 Using Nanobank

An old English saying holds that "The proof of the pudding is in the eating," and so the value of Nanobank (and COMETS) can only be judged by the research that it enables. A number of papers in this special issue of *Annales* make a down payment on that program, and a substantially larger number of research projects by users of Nanobank (and now COMETS) are underway. Here we present some simple uses of Nanobank by way of illustration and suggestion of its capabilities in providing data for more extensive research projects.

It is extremely difficult if not impossible to measure firm entry in any given country, much less comparably across countries. Darby and Zucker (2005) demonstrated that first appearance as author's address on a nano-article or assignee on a nano-patent served as a useful measure of entry for nanotechnology companies. Zucker and Darby (2014) confirmed this for across the range of sciences and high-technology industries and specifically found no important difference in the results for firm entry whether this proxy or a directory-and-web-based enumeration of nano-firm entry was used. In a series of articles reviewed in Zucker and Darby (2009, 2014) the senior authors and their coauthors have shown that the very top "star" scientists are key determinants of where and when firms with related high-technologies enter and which firms are most successful.

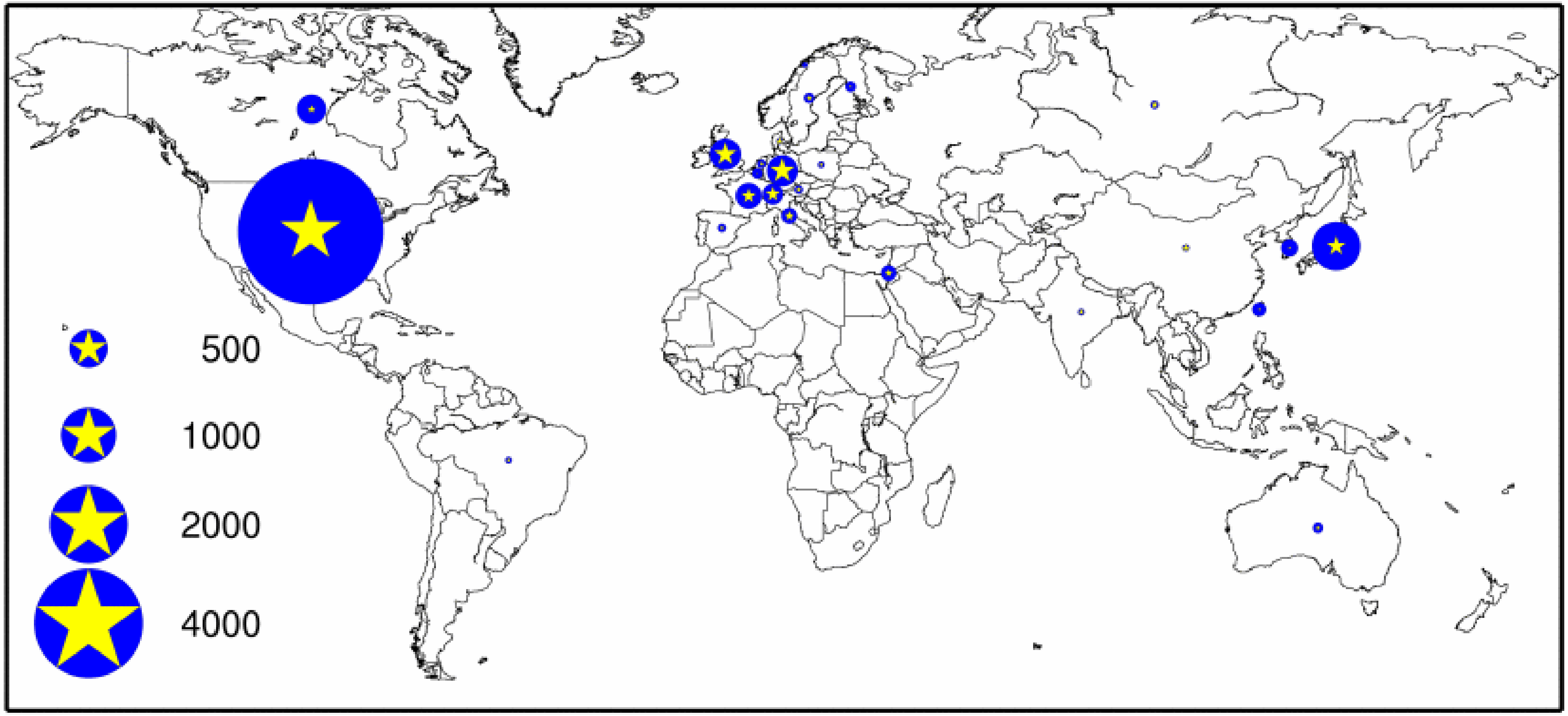
Figures 2 and 3 show how the data in Nanobank can be used to measure firm entry across regions (B.E.A. functional economic areas) in the United States and across countries in the world where the cumulative number of firm entries over 1981-2004 (the bulk since 1990) are indicated by

Figure 2. Locations of US Star Nano-Scientists (★) and Cumulative Nano-Firm Entries (●) by Region 1981-2004



Note: An animated version of this figure showing star locations and firm entries by year and a comparable animation for non-nano stars and non-nano-firm entries is at http://www.nanoconnection.net/research/results/2006/stars_firms_us.php

Figure 3. Locations of World Star Nano-Scientists (★) and Cumulative Nano-Firm Entries (●) by Country 1981-2004



Note: An animated version of this figure showing star locations and firm entries by year and a comparable animation for non-nano stars and non-nano-firm entries is at http://www.nanobank.org/research/results/stars_firms_world.php the size of the circles.[4] We use the ISI Highly Cited authors to define nano-stars for the purposes of these maps. The high correlation between the number of stars (indicated by the size of the stars) and then number of entries is even more striking in the animated maps noted at the bottom of each figure. Zucker and Darby (2009, 2014) report rigorous multivariate statistical tests which confirm the impression from the maps.

3 Constructing the COMETS Database

Construction of the COMETS data base under the STAR and COMETS project 2007-2015 builds on the methodology and groundwork of Nanobank. However, the goals have expanded considerably to cover all sciences and high-technologies and ultimately to span the national innovation system from government funding and policies through scientific advance and industrial formation and transformation. Further, policy change at a key vendor due to a change in ownership between the starts of the Nanobank and STAR and COMETS projects limits availability of articles data except in aggregated analysis data sets to a limited number of on-site users at NBER and UCLA. Nonetheless, with this exception for post-2005 articles data, COMETS data can be used in lieu of Nanobank data as we include information indicating which patents and grants – and other records as they are added – are identified as being relevant to nano-scale science and engineering, and the particular methods used to make the identifications (keywords, probabilistic, authoritative).

The Kauffman Foundation encouraged early use of the data through a COMETS Travel Grants Program that supported the presentation of COMETS data based empirical research at conferences through direct grants to users. Numerous papers using COMETS, Nanobank, or both have since been presented, and many published, spreading the word in the most way about the value of these databases.

The conceptual structure of COMETS is illustrated in Figure 4. The ovals represent the major actors in the national innovation system and the connecting hypothesized represent flows of resources, knowledge, and/or innovation among them. Identified data sources for which we have acquired rights to use data are indicated in the ovals for which they are most relevant. Type-face codes for the data sources indicate whether the data are available in COMETS 2.0 (the latest version as this article goes to press), are planned for future releases (after beta-testing in COMETSbeta and COMETSandSTARS), or due to contractual restrictions imposed by vendors will be available only to on-site COMETSandSTARS users at NBER and/or UCLA.[5] For each added legacy data set, considerable time and effort is required for parsing legacy data sets into usable fields; cleaning the data for both vendors' and our own processing errors; matching organizations and scientists with those currently in the database and creating new IDs for new cases of each; and managing the beta test and responding to users comments and corrections. As a result, how far the Zucker-Darby team

[4] Figure 3 displays data for only the top-25 science and technology countries in the world, but they have essentially all the nano-firm entry and nano-stars in the world.

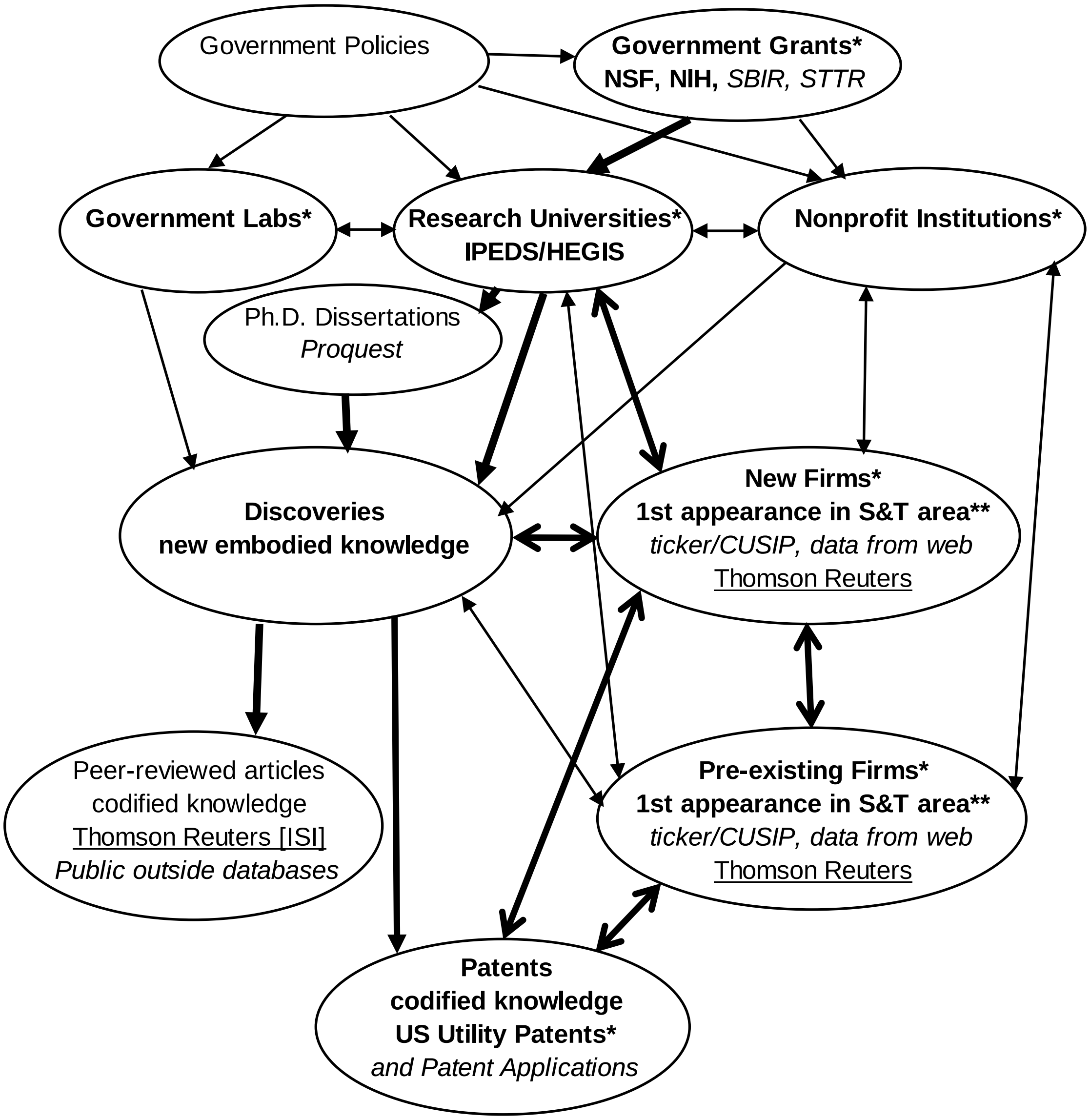
[5] Figure 4 is updated in press to the latest version in order to provide readers the most up-to-date status. Changes from COMETS and for future upgrades are documented as they are made at the COMETS, COMETSbeta, and COMETSandSTARS websites and on-site codebooks and other records.

can go in completing the build-out of COMETS – let alone developing data sources for the currently empty ovals – is dependent on the availability of follow-on funding.

Section 3.1 describes the first version of COMETS – COMETS 1.0, laying out its contents in some detail. Section 3.2 documents the flags used to indicate five (six counting nanotechnology) major S&T areas cutting across the records from funding and basic discoveries to patented

Figure 4. Architecture and Contents of the COMETS Database

Architecture and Contents of Kauffman.org/COMETS


Government Policies
Government Grants* NSF, NIH, SBIR, STTR
Government Labs*
Research Universities* IPEDS/HEGIS
Nonprofit Institutions*
Ph.D. Dissertations Proquest
Discoveries new embodied knowledge
New Firms* 1st appearance in S&T area** ticker/CUSIP, data from web Thomson Reuters
Peer-reviewed articles codified knowledge Thomson Reuters [ISI] Public outside databases
Pre-existing Firms* 1st appearance in S&T area** ticker/CUSIP, data from web Thomson Reuters
Patents codified knowledge US Utility Patents* and Patent Applications

Status as of version COMETS 2.0, February 12, 2014:
Bold = available now *Italic = coming in future versions* Underlined = COMETSandSTARS only

* Organization ID, matched across data clusters/types. Person name as it appears in data - inventor, principal investigator, author (NanoBank.org).

** First appearance in the patent or grant databases identifies birth in a specific S&T area.

technologies and industrial classifications. The procedure we are currently using to obtain unique IDs associated with an individual whenever they appear in any of the constituent legacy databases is laid out in Section 3.3.

3.1 COMETS 1.0

COMETS version 1.0 – the initial release at the Kauffman Foundation website – integrates the US patents data with NIH and NSF grants data. Comments from the 100-plus beta-testers indicate that even those just interested, say, in using the patent data find the parsed and matched data much preferable to using data available from the US Patent and Trademark Office. The COMETS 1.0 database includes 3,911,920 US patents with grant dates from 1976 to 2010. The government grants data includes 418,054 NIH grants and 345,574 NSF grants from 1972 to 2010. A description of the contents of the COMETS 1.0 database is included in Table 4. As with the corresponding tables for Nanobank, a careful perusal of Table 4 will reward the reader with an understanding of the large number of variables included and even larger number which can be constructed using the information in COMETS.

3.2 Science and Technology Areas in COMETS

In our work on biotechnology, it was possible to track a relatively narrowly defined body of knowledge from its origins (largely in universities), to development of inventions represented by patents, to commercial applications in firms and ultimately into goods and service in the market place. Nanobank aims to define a similarly relative narrow but even more broadly interdisciplinary set of articles, patents, and firms with the affiliation and/or location of individual participants identified so far as possible. It is natural to want to compare activities in nanotechnology (or biotechnology) with those in other science and technology (S&T) areas, but in attempting to do so we learned that it was generally more difficult to find narrowly defined areas of science (categorizing articles and doctoral programs) that correspond to narrowly defined areas of technology (categorizing by patent classes) that correspond to narrowly defined areas of industry (categorizing by governmental or financial market definitions of industry).

In Darby and Zucker (1999) and Zucker and Darby (1999) we developed and detailed a concordance across five science and engineering areas, technological areas, and industrial applications for analyses that spanned scientific articles, patents, and university doctoral-programs data from the National Research Council (1995): Biology, Chemistry & Medicine; Computing & Information Technology; Semiconductors, Integrated Circuits & Superconductors; Other Sciences; and Other Engineering. We were unable to find finer breakdowns that did not require data in greater detail than existed in one or more of these sources. Our experience since has been that this concordance is generally useful for a number of analytical purposes and we make it available in COMETS for others who might be inclined to use it in their work. In our own use of these areas, we create a sixth S&T area of nanotechnology by subtracting the records flagged as nano-S&T related from the S&T areas in which they would otherwise appear. Detailed concordances are posted at http://www.nanobank.org/downloads.php.

By way of example, in research described in Zucker and Darby (2014) we applied these categories, with articles and firms based upon the Nanobank technologies subtracted to form a sixth specific Nanoscale Science & Technology area. That analysis showed that firms in all six areas were more likely to be founded in countries or U.S. regions when and where top “star” scientists and engineers for the given S&T area were resident. In this case both surprising similarities and interesting variations in patterns of firm birth and star migration were observed. We hope that they will prove equally useful for other purposes beyond their origin.

Table 4. COMETS Data Description as of February 12, 2014

Table Name	Column Name	Col. Type	Column Description	Reference Sheet
SECTION 1 : Patents				
patents	patent_id	integer	patent number	
	num_claims	integer	number of claims	
	grant_date	date	grant date	
	app_num	character	application number	
	app_date	date	application date	
	patent_title	character	patent title	
patent_citations	cite_from_patent_id	integer	patent number of citing patent	
	cite_from_patent_gyr	integer	grant year of citing patent	
	cite_to_patent_id	integer	patent number of cited patent	
	cite_to_patent_gyr	integer	grant year of cited patent	
patent_cite_counts	patent_id	integer	patent number	
	grant_year	integer	grant year	
	citations	integer	# of patents granted this year that cite this patent	
patent_int_classes	patent_id	integer	patent number	
	pos	integer	order of appearance for this class	
	intl_class	character	international patent class	
patent_us_classes	patent_id	integer	patent number	
	pos	integer	order of appearance for this class	
	us_class	character	US patent class	
patent_abstracts	patent_id	integer	patent number	
	patent_title	character	patent title	
	patent_abstract	character	patent abstract	
patent_assignees	patent_id	integer	patent number	
	pos	integer	order of appearance for this assignee	
	org_name	character	name of university, company, institution, etc.	
	org_id	character	alphanumeric code specific to each organization	Org Codes Info
	org_type	character	organization type	Org Codes Info
	org_norm_name	character	normalized name	Org Codes Info
	city	character	city	
	state	character	state	
	country	character	country	
	country_code	integer	ISO country code	ISO Country Info
	bea_code	integer	BEA code	BEA Info
	num_claims	integer	number of claims	
	grant_date	date	grant date	
	app_num	character	application number	
	app_date	date	application date	

Table 4. COMETS Data Description as of February 12, 2014 (continued)

Table Name	Column Name	Col. Type	Column Description	Reference Sheet
patent_inventors	patent_id	integer	patent number	
	pos	integer	order of appearance for this inventor	
	last_name	character	last name	
	first_name	character	first name	
	middle_name	character	middle name	
	suffix	character	name suffix	
	street	character	street address	
	city	character	city	
	state	character	state	
	postal_code	character	postal code	
	country	character	country	
	country_code	integer	ISO country code	ISO Country Info
	bea_code	integer	BEA code	BEA Info
	num_claims	integer	number of claims	
	grant_date	date	grant date	
	app_num	character	application number	
	app_date	date	application date	
patent_zd_cats	patent_id	integer	patent number	
	zd	character	Zucker-Darby Science and Technology Area Category	ZD Categories
	weight	decimal	fractional category weight (0.0 to 1.0)	
patent_nano	patent_id	integer	patent number	
	is_nano	integer	1 if identified as nano-related, 0 otherwise	NANO Identification
	is_nano_bool	integer	1 if identified as nano-related by boolean method	
	is_nano_prob1	integer	1 if identified as nano-related by probabilistic method #1	
	is_nano_prob2	integer	1 if identified as nano-related by probabilistic method #2	
	is_nano_auth	integer	1 if identified as nano-related by an authoritative source	
SECTION 2 : Grants				
grants	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	fiscal_year	integer	fiscal year	
	start_date	date	start date	
	end_date	date	end date	
	last_amend_date	date	last amendment date	
	instrument	character	award instrument	
	amount	integer	award amount	
	grant_title	character	grant title	
grantee_orgs	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	org_name	character	name of university, company, institution, etc.	
	org_id	character	alphanumeric code specific to each organization	Org Codes Info
	org_type	character	organization type	Org Codes Info
	org_norm_name	character	normalized name	Org Codes Info
	street	character	street address	
	city	character	city	
	state	character	state	
	postal_code	character	postal code	
	country	character	country	
	country_code	character	ISO country code	ISO Country Info
	bea_code	integer	BEA code	BEA Info
grant_pis	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	last_name	character	PI last name	
	first_name	character	PI first name	
	middle_name	character	PI middle name	

Table 4. COMETS Data Description as of February 12, 2014 (continued)

Table Name	Column Name	Col. Type	Column Description	Reference Sheet
patent_inventors	patent_id	integer	patent number	
	pos	integer	order of appearance for this inventor	
	last_name	character	last name	
	first_name	character	first name	
	middle_name	character	middle name	
	suffix	character	name suffix	
	street	character	street address	
	city	character	city	
	state	character	state	
	postal_code	character	postal code	
	country	character	country	
	country_code	integer	ISO country code	ISO Country Info
	bea_code	integer	BEA code	BEA Info
	num_claims	integer	number of claims	
	grant_date	date	grant date	
	app_num	character	application number	
	app_date	date	application date	
patent_zd_cats	patent_id	integer	patent number	
	zd	character	Zucker-Darby Science and Technology Area Category	ZD Categories
	weight	decimal	fractional category weight (0.0 to 1.0)	
patent_nano	patent_id	integer	patent number	
	is_nano	integer	1 if identified as nano-related, 0 otherwise	NANO Identification
	is_nano_bool	integer	1 if identified as nano-related by boolean method	
	is_nano_prob1	integer	1 if identified as nano-related by probabilistic method #1	
	is_nano_prob2	integer	1 if identified as nano-related by probabilistic method #2	
	is_nano_auth	integer	1 if identified as nano-related by an authoritative source	
SECTION 2 : Grants				
grants	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	fiscal_year	integer	fiscal year	
	start_date	date	start date	
	end_date	date	end date	
	last_amend_date	date	last amendment date	
	instrument	character	award instrument	
	amount	integer	award amount	
	grant_title	character	grant title	
grantee_orgs	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	org_name	character	name of university, company, institution, etc.	
	org_id	character	alphanumeric code specific to each organization	Org Codes Info
	org_type	character	organization type	Org Codes Info
	org_norm_name	character	normalized name	Org Codes Info
	street	character	street address	
	city	character	city	
	state	character	state	
	postal_code	character	postal code	
	country	character	country	
	country_code	character	ISO country code	ISO Country Info
	bea_code	integer	BEA code	BEA Info
grant_pis	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	last_name	character	PI last name	
	first_name	character	PI first name	
	middle_name	character	PI middle name	

Table 4. COMETS Data Description as of February 12, 2014 (continued)

Table Name	Column Name	Col. Type	Column Description	Reference Sheet
grant_co_pis	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	pos	integer	order of appearance of this co-PI	
	last_name	character	co-PI last name	
	first_name	character	co-PI first name	
	middle_name	character	co-PI middle name	
grant_abstracts	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	grant_title	character	grant title	
	grant_abstract	character	grant abstract	
grant_zd_cats	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	zd	character	Zucker-Darby Science and Technology Area Category	ZD Categories
	weight	decimal	fractional category weight (0.0 to 1.0)	
grant_nano	grant_agency	character	granting agency	
	grant_num	character	agency's grant number	
	is_nano	integer	1 if identified as nano-related, 0 otherwise	NANO Identification
	is_nano_bool	integer	1 if identified as nano-related by boolean method	
	is_nano_prob1	integer	1 if identified as nano-related by probabilistic method #1	
	is_nano_prob2	integer	1 if identified as nano-related by probabilistic method #2	
	is_nano_auth	integer	1 if identified as nano-related by an authoritative source	
SECTION 2.1 : NSF-Specific Grant Information				
grant_nsf	grant_num	character	agency's grant number	
	prog_manager	character	program manager	
	directorate	character	NSF directorate	
grant_nsf_programs	grant_num	character	agency's grant number	
	pos	integer	order of appearance of this program	
	program	character	NSF program name	
grant_nsf_fields	grant_num	character	agency's grant number	
	pos	integer	order of appearance of this field of application	
	field	character	NSF field of application	
	field_code	character	code for this field of application	
SECTION 2.2 : NIH-Specific Grant Information				
grant_nih	grant_num	character	agency's grant number	
	nih_icd	character	NIH institute, center, or division	
	nih_irg	character	NIH initial review group	
grant_nih_tags	grant_num	character	agency's grant number	
	pos	integer	order of appearance of this tag	
	tag	character	NIH descriptive tag	
grant_nih_core_proj_nums	grant_num	character	agency's grant number	
	nih_core_proj_num	character	NIH core project number	

Table 4. COMETS Data Description as of February 12, 2014 (continued)

Table Name	Column Name	Col. Type	Column Description	Reference Sheet
SECTION 3 : Universities				
univ_org_info	unitid	character	IPEDS identifier	
	org_id	character	alphanumeric code specific to each organization	
	org_type	character	organization type	
univ_institutional_info	unitid	character	IPEDS identifier	
	instnm	character	institution name from IPEDS	
	city	character	city	
	stabbr	character	state	
	zip	character	zip code	
	public_private_status	character	type of institution (private or public)	
	yr	integer	year	
	bea_2005	character	BEA code	BEA_Info
univ_enrollments	unitid	character	IPEDS identifier	
	total_ft_ug	integer	Total full time undergraduate	
	total_ft_grad	integer	Total full time graduate	
	total_pt_ug	integer	Total part time undergraduate	
	total_pt_grad	integer	Total part time graduate	
	total_enr	integer	Total undergraduate and graduate enrollment	
	yr	integer	year	
univ_degrees	unitid	character	IPEDS identifier	
	awlevel	character	award level	
	degrees	integer	total degrees awarded	
	yr	integer	year	
	zd	character	Zucker-Darby Science and Technology Area	
univ_systems	system_org_id	character	University system org_id	
	system_name	character	University system name	
	campus_org_id	character	University campus org_id	
	campus_name	character	University campus name	
univ_faculty_extended	unitid	character	IPEDS identifier	
	yr	integer	year	
	prof_men_9_10_total	integer	Professors, 9/10 month contract, men	
	assoc_prof_men_9_10_total	integer	Associate professors, 9/10 month contract, men	
	assist_prof_men_9_10_total	integer	Assistant professors, 9/10 month contract, men	
	instruct_men_9_10_total	integer	Instructors, 9/10 month contract, men	
	lec_men_9_10_total	integer	Lecturers, 9/10 month contract, men	
	no_rank_men_9_10_total	integer	No rank, 9/10 month contract, men	
	total_fac_men_9_10	integer	Total faculty, 9/10 month contract, men	
	prof_women_9_10_total	integer	Professors, 9/10 month contract, women	
	assoc_prof_women_9_10_total	integer	Associate professors, 9/10 month contract, women	
	assist_prof_women_9_10_total	integer	Assistant professors, 9/10 month contract, women	
	instruct_women_9_10_total	integer	Instructors, 9/10 month contract, women	
	lec_women_9_10_total	integer	Lecturers, 9/10 month contract, women	
	no_rank_women_9_10_total	integer	No rank, 9/10 month contract, women	
	total_fac_women_9_10	integer	Total faculty, 9/10 month contract, women	
	total_fac_9_10	integer	Total faculty, 9/10 month contract	
	prof_men_11_12_total	integer	Professors, 11/12 month contract, men	
	assoc_prof_men_11_12_total	integer	Associate professors, 11/12 month contract, men	
	assist_prof_men_11_12_total	integer	Assistant professors, 11/12 month contract, men	
	instruct_men_11_12_total	integer	Instructors, 11/12 month contract, men	
	lec_men_11_12_total	integer	Lecturers, 11/12 month contract, men	
	no_rank_men_11_12_total	integer	No rank, 11/12 month contract, men	
	total_fac_men_11_12	integer	Total faculty, 11/12 month contract, men	
	prof_women_11_12_total	integer	Professors, 11/12 month contract, women	
	assoc_prof_women_11_12_total	integer	Associate professors, 11/12 month contract, women	
	assist_prof_women_11_12_total	integer	Assistant professors, 11/12 month contract, women	
	instruct_women_11_12_total	integer	Instructors, 11/12 month contract, women	
	lec_women_11_12_total	integer	Lecturers, 11/12 month contract, women	
	no_rank_women_11_12_total	integer	No rank, 11/12 month contract, women	
	total_fac_women_11_12	integer	Total faculty, 11/12 month contract, women	
	total_fac_11_12	integer	Total faculty, 11/12 month contract	
	total_other_fac	integer	Total other faculty	

Table 4. COMETS Data Description as of February 12, 2014 (concluded)

Table Name	Column Name	Col. Type	Column Description	Reference Sheet
univ_system patents	org_id	character	ZD org_id	
	yr	integer	year	
	zd	character	ZD area	
	num_patents	integer	number of patents	
univ_student demographics	unitid	character	IPEDS identifier	
	yr	integer	year	
	non_res_alien_men	integer	total non resident alien men	
	non_res_alien_women	integer	total non resident alien women	
	black_non_hisp_men	integer	total black non hispanic men	
	black_non_hisp_women	integer	total black non hispanic women	
	amer_ind_alask_nat_men	integer	total american indian and alaskan native men	
	amer_ind_alask_nat_women	integer	total american indian and alaskan native women	
	asian_pacisl_men	integer	total asian and pacific islander men	
	asian_pacisl_women	integer	total asian and pacific islander women	
	hisp_men	integer	total hispanic men	
	hisp_women	integer	total hispanic women	
	white_non_hisp_men	integer	total white non hispanic men	
	white_non_hisp_women	integer	total white non hispanic women	
	race_ethnicity_unk_men	integer	total race/ethnicity unknown men	
	race_ethnicity_unk_women	integer	total race/ethnicity unknown women	
	total_men	integer	total men	
	total_women	integer	total women	

The concordance as posted is organized in three tables for articles, patents, and NRC (1995) doctoral programs. Each of these tables contains a list of document categorizations and the corresponding Zucker-Darby category codes and descriptions. The categorizations for articles are the journal categories assigned by the ISI Web of Knowledge. The patent categorizations are the International Patent Classifications assigned by the World Intellectual Property Organization. The categorizations for NRC doctoral programs are the NRC standard doctoral programs. Corresponding tables for industries are being prepared and will be posted in the near future. Please contact us if you want to be notified as soon as they are available and the number of such requests will guide the priorities for our available time for building new data series.

3.3 Person Matching

The greatest challenge in building both Nanobank and COMETS has proven to be person matching or disambiguating tens of millions of observations of individuals' names down to millions of unique individuals acting variously as inventors, principal investigators, authors, entrepreneurs, chief scientists, and other guises. There are in fact a number of active scientists with exactly identical names – some family names are common and certain combinations of family and given names are more appealing to parents than others. The substantive problems are: (a) In all the legacy databases (patents, research articles, grants, the various financial databases) there is no attempt to assign a unique identifier used each time a certain individual appears.[6] (b) A given individual's names may appear differently depending on the conventions applied by the particular person or institution inputting the data (e.g., a patent attorney or journal editor) or to changing circumstances or habits of the individual (e.g., marriage, dropping a middle name with increased fame). (c) The research article

[6] There is an internal effort to use unique identifiers at the federal granting agencies, but that effort is not reflected as of this date in the available databases.

data until very recent years gave only family name and initials for given names and associated the addresses listed on the article only in the case of the corresponding author.[7] (d) The other-than-name information known for each individual observation varies even within legacy databases and more so across legacy databases – e.g. work addresses in grants, articles, and financial data versus residence addresses (usually with missing street address) in patents. (e) Scalability of matching methods becomes an issue as the number of calculations and probability comparisons rises exponentially with the number of unique observations and hence possible matches. (f) Using information about an individual gleaned from other legacy databases can improve the quality of the matches in a given legacy database, but this implies an inherent iteration which multiplies the scalability issues. (g) Data quality is hard to assess and reservation of known individuals for quality checking means the probability calculations in the match are less accurate than if the reserved data had been used in estimation.

The methodology we use for person matching can be outlined as follows.

1. First we either locate or build a learning set of thousands of individuals for whom we have or can obtain essentially complete data across the main legacy databases.
2. We then simplify the problem by considering only cases for which the family name and first initial are the same as possible matches. This means that we will never match misspelled family names or errors in first initials, but it makes the problem computationally tractable.[8]
3. Next for each legacy database we collect all possible matches to the individuals in the learning set and use the listed names in each observation plus such collateral information as address match, other individuals on the same record, keywords, S&T field for record, journal match, to calculate probability estimators based on the learning set for each legacy database and for across database pairs.
4. The match begins within each legacy database by imposing some definite matches with probability 1 (e.g., for authors or inventors self-citing prior articles or patents and for continuing groups [half or more the same] of co-authors or co-inventors).
5. Next probabilities are computed for every possible remaining pairwise match for each last name first initial combination. Those pairs with probabilities above a selected threshold are declared matched, starting with the pair with the highest probability and then going on to the highest among the remaining unpaired records until no remaining pairs meet the threshold.
6. Using all the information in each group (initially pairs) the probability that other groups or unmatched records is a match is computed and those with a probability above a second (lower) threshold are declared matched. This process is iterated until there are no remaining

[7] The addresses (almost always a work address) given in the journal are all listed, but only one of these is associated with a particular author designated as the corresponding author. We know the address(es) associated with an author only if that author is the corresponding author, a sole author, or one of several authors on an article with only one listed address in a year in which the journal lists multiple addresses on other articles. Note that even for a corresponding author, we only know one address of possibly several addresses even if she lists dual affiliations unless she is the sole author.

[8] Person matching is done off-line on the CISTCP cluster running 32 Sun processors in parallel as required by vendors' licensing terms. Nonetheless, a full run with a single set of probability parameters takes weeks, not days.

matches meeting the second higher grouping threshold. The second threshold used for matching between groups of observations is lower than the first threshold used for matching between single observations because there is usually more information available when considering groups of observations. The first threshold is higher to avoid creating false matches when less information is available and there are more instances of missing information. If a true match is missed due to this higher threshold, there will still be additional opportunities to create a match in the second pass for group matching.

7. Next all information in each group (including groups of size 1) created within each legacy database is used to compute probabilities of matching with every group in other legacy databases. The higher grouping threshold is again applied to create cross-database matches. This process iterates until no further matches meet the threshold.
8. Unique ID numbers are assigned to each of the groups (including the groups of size one which are treated as single appearances by a unique individual).

In mid-August 2011 we were very close to having a full match to test against known matches for Type I and II errors in articles and patents. We learned from that test that matching across data sets with very different information can greatly enrich knowledge about individual scientists, but we see that active work/analysis with these data are necessary for improvement of the match and removal of systematic errors, such as systematic changes in original data cleaning, as in ISI where compound last names (e.g., Burne-Jones) are initially separated into two words separated by a space (Burne Jones) and later run together with no hyphen (BurneJones). Hence, we are now working to further improve matching procedures, beginning with patent data. Person IDs for those databases which permit it will initially be available in COMETSandSTARS only. We will test the match quality there while monitoring the experiences of the Fleming group with full public release of person level identifiers in patents (Li et al. 2014 in press). We expect then to add individual IDs for persons wherever they appear as inventors and authors in the patent data, NSF and NIH grant data, and a prototype open-source articles database, first in COMETSbeta, and then into the COMETS and Nanobank sites.

4 Conclusions

COMETS can be seen as a work in progress, and it clearly is. The Zucker-Darby team has an ambitious agenda to complete processing, testing, and adding to the COMETS files important legacy databases which will deepen the community's ability to develop tested knowledge on the processes of knowledge formation and use in discovery, innovation, technological progress, and economic growth. After the first three months, more than 130 beta-test users are using the data so far provided in COMETS 1.0. Science policy, economic growth and the nation will all profit from their efforts. If funding is available to complete the COMETS build-out, among our next steps are re-engineering of data on public firms primarily drawn directly from edgar.gov and other government public sources, and a re-engineering of public science sources, including Google, to develop new data on links between academic scientists and firms which can be posted on the public website. We will include an updated and extended Nanobank database including public article data within COMETS, allowing research on this important new S&T area up through 2012 instead of the current cut-off at 2005. The authors hope that further extensions and enhancements will be undertaken by a permanent institutional home charged with maintaining and developing Connecting Outcome Measures in Entrepreneurship, Technology and Science. We believe that the best way to ensure continued national investment in the scientific seed corn of scientific, economic, and social policy is by documenting carefully for the public and their representatives the impressive payoffs earned on their investments.

In conclusion, the Nanobank and COMETS databases provide very flexible sources of micro data for serious research on all S&T areas, on nanotechnology as a special case, and on a variety of issues in science of science and technology. Researchers are most welcome to try it for themselves.

REFERENCES

[1] Darby, Michael R., Zucker, Lynne G. (1999) *California's Science Base: Size, Quality and Productivity*, Sacramento, CA: California Council on Science and Technology.

[2] Darby, Michael R., Zucker, Lynne G. (2005) “Grilichesian Breakthroughs: Inventions of Methods of Inventing in Nanotechnology and Biotechnology,” *Annales d’Economie et Statistique*, July/December, 79/80, 143-164.

[3] Darby, Michael R., Zucker, Lynne G. (2007) “Real Effects of Knowledge Capital on Going Public and Market Valuation,” in Naomi Lamoreaux and Kenneth Sokoloff, eds., *Financing Innovation in the United States, 1870 to the Present*, Cambridge, MA: MIT Press.

[4] Li, Guan-Chen, Lai, Ronald, D'Amour, Alexander, Doolin, David M., Sun, Ye, Torvik, Velte I., Yu, Amy Z., Fleming, Lee. (2014 in press) “Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010),” *Research Policy*.

[5] Liebeskind, Julia Porter, Oliver, Amalya, Zucker, Lynne G., Brewer, Marilynn B. (1996) “Social Networks, Learning, and Flexibility: Sourcing Scientific Knowledge in New Biotechnology Firms.” *Organization Science*, July/August, 7, 428-443.

[6] National Research Council (1995) *Research-Doctorate Programs in the United States: Data Set,* machine-readable data base, Washington, US: National Academy Press.

[7] Zucker, Lynne G., Darby, Michael R. (1996) “Star Scientists and Institutional Transformation: Patterns of Invention and Innovation in the Formation of the Biotechnology Industry,” *Proceedings of the National Academy of Sciences*, Nov. 12, 93(23), 12709-12716.

[8] Zucker, Lynne G., Darby, Michael R. (1997) “Present at the Biotechnological Revolution: Transformation of Technical Identity for a Large Incumbent Pharmaceutical Firm.” *Research Policy,* 26, 429-446.

[9] Zucker, Lynne G., Darby, Michael R. (1999) *California's Inventive Activity: Patent Indicators of Quantity, Quality, and Organizational Origins,* Sacramento, US: California Council on Science and Technology.

[10] Zucker, Lynne G., Darby, Michael R. (2009) “Star Scientists, Innovation and Regional and National Immigration,” in David B. Audretsch, Robert E. Litan, and Robert J. Strom, eds., *Entrepreneurship and Openness: Theory and Evidence,* volume 2 in the series *Industrial*

Dynamics, Entrepreneurship and Innovation, Cheltenham, UK, and Northampton, MA: Edward Elgar.

[11] Zucker, Lynne G., Darby, Michael R. (2011) "Legacy and New Databases for Linking Innovation to Impact," in Kaye Husbands Fealing, Julia Lane, John H. Marburger III, Stephanie Shipp, eds., *The Science of Science Policy: A Handbook*, Palo Alto, CA: Stanford University Press.

[12] Zucker, Lynne G., Darby, Michael R. (2014) "Movement of Star Scientists and Engineers and High-Tech Firm Entry," *Annals of Economics and Statistics (Annales d'Economie et Statistique)*, this issue.

[13] Zucker, Lynne G., Darby, Michael R., Armstrong, Jeff (1998) "Geographically Localized Knowledge: Spillovers or Markets?" *Economic Inquiry*, January, 36(1), 65-86.

[14] Zucker, Lynne G., Darby, Michael R., Armstrong, Jeff (2002) "Commercializing Knowledge: University Science, Knowledge Capture, and Firm Performance in Biotechnology." *Management Science*, January, 48(1), 138-153.

[15] Zucker, Lynne G., Darby, Michael R., Brewer, Marilynn B. (1998) "Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises," *American Economic Review*, March, 88(1), 290-306.