Keith G Jeffery

Jan Dvořák

(editors)

e-Infrastructures for Research and Innovation

Linking Information Systems to Improve Scientific Knowledge Production

Proceedings

of the

11th International Conference

on

Current Research Information Systems

CRIS 2012

Prague, Czech Republic

June 6th–9th, 2012

Creating a Data Infrastructure for Tracking Knowledge Flow

Laurel L. Haak[a], David Baker[b], Matt A. Probus[c]

[a] ORCID, [b] CASRAI, [c] Thomson Reuters

Summary

To understand how research and development leads to creation of knowledge and then to track the impact of that knowledge requires a comprehensive model of the research ecosystem that incorporates inputs, outputs, activities, and external factors, and the data to support longitudinal and network analysis. To date, most research has focused on those activities and outputs that are readily accessible, including publication output and follow-on citations, and patents and patent citations. While these outputs are robust and can be normalized by field of research, additional data are needed. Moreover, efforts to assemble systematic information on researchers, including their biographic information, institutions, support, and networks, are in a fledgling stage. We discuss requirements around data linkages, data standards, and data privacy in creating a distributed data infrastructure to support quantitative analysis of the research workforce.

1 Deriving linkages between data sources

Understanding the key components that support or inhibit the flow of knowledge in the research and development (R&D) ecosystem is critical for effective formulation of research management policy, strategy, and execution. Knowledge flows underlie identification of experts, assessment of capabilities, workforce planning, technology transfer, grants portfolio management, and are necessary part of program outcomes evaluation.

The data necessary to study the research workforce and associated knowledge flow dynamics are currently scattered across sources and formats that for the most part have not been linked. Information on knowledge flow can be as varied as maps of the movement of people in a research space (who talked to whom and where), interviews with program participants (who worked with whom on what and why), publications, patents, computer code, financial records, and flight logs.

We discuss creation of an information infrastructure to connect data about R&D knowledge flow. This infrastructure would establish a dictionary of variables and standards for data exchange between platforms that, at the end of the day, will allow the R&D community to perform analysis and modeling of knowledge flows and support data-driven research policy decisions.

2 ScienceWire: Showing the promise of linking

A promising platform in this area is ScienceWire (Figure 1), a large-scale linked data infrastructure, the goal of which is to provide a common data infrastructure and analysis procedures for the research community to catalyze research on innovation. ScienceWire ingests and transforms raw data to create document-type catalogs. Currently, ScienceWire consists of catalogs housing publication data, patent data, and awarded grant data. These catalogs integrate both syntactic and semantic interoperability to provide a standard structure and format for storing similar documents. This normalization allows the consolidation of metadata onto one document record. Coupling text mining with enriched records has supported the creation of linkages between documents within a catalog (such as co-author networks) and between catalogs (such as grant funding impact on patenting activity and drug development).

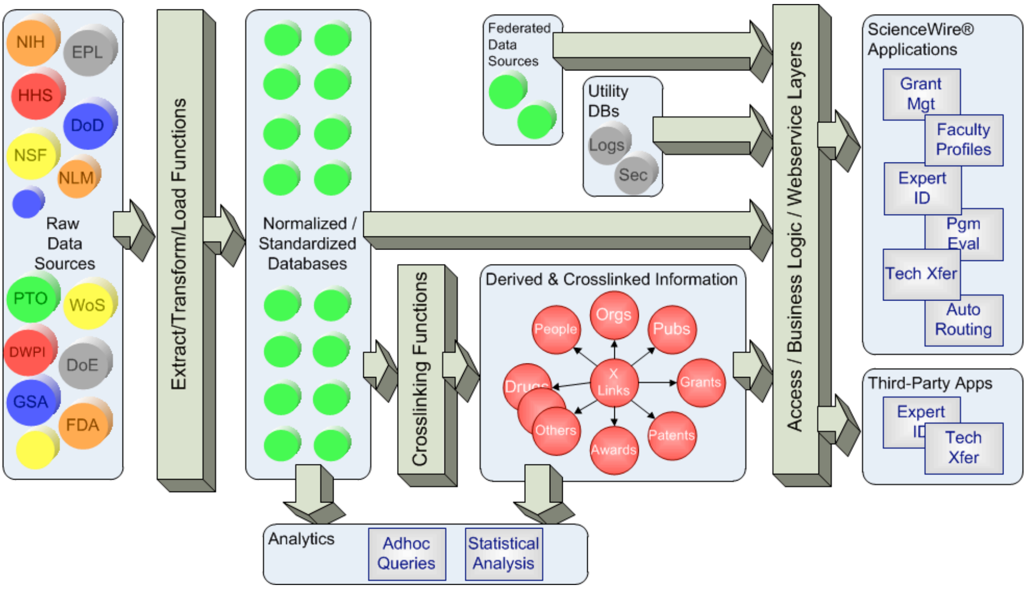

Figure 1: ScienceWire Architecture

One example of the usefulness of this approach is a project carried out to link research funding to new drug approvals using data from all three catalogs (Wright et al. 2011). This study proved the feasibility of using metadata on grant support extracted from patents to automate the linkage of US National Institutes of Health (NIH) grants → US Food and Drug Administration (FDA) approved drugs. Three methods to link FDA-approved drugs to NIH grants were employed. All methods began by linking FDA approved drugs directly to the patents acknowledged on the New Drug Application (NDA). Method 1 used direct linkage to projects through an NDA patent → grant acknowledgement. Method 2 used indirect linkage through a patent → patent citation, and Method 3 used linkage through a patent →publication citation. Method 1 was found by subject matter experts to be 100% reliable in identifying direct and substantive contributions of projects to FDA-approved drugs, while Methods 2 & 3 were found to accurately capture all relevant contributions, but required additional review to validate the strength of the linkage.

3 Open Standards: The keys to distributed linking

It is impossible to track information flow if the information has conflicting meanings, structures, identifiers and formats across systems. That no one central infrastructure is likely to capture all the data needed for effective knowledge flow analysis means that an effort must be made to make a network of infrastructures interoperable, regardless of local storage formats. It is important to integrate this effort with others, such as the disambiguation of USPTO patent inventors (Lai et al. 2010) or the US Federal STAR METRICS initiative (Lane 2010), and also to broadly leverage existing initiatives, such as the CrossRef DOI, ORCID, CERIF syntax, and the CASRAI dictionary. Accessible, interoperable data will also make it possible for the community to dynamically refine the underlying data.

Ongoing efforts to standardize vocabularies on research information include the Frascati manual (OECD 2007) and the UNESCO SPINES thesaurus (UNESCO 1976). The European Union has adopted the Common European Research Information Format (CERIF) structure for research management systems. Recently, partnerships have been forming between key organizations (euroCRIS, CASRAI, VIVO, JISC, and others) and this emerging family of integrated international non-profit organizations are forming one of a number of key components in a global approach. Among the benefits of this collaboration will be shared nomenclatures and crosswalks between existing ontologies.

As one example in this ecosystem, CASRAI provides a standard dictionary of meaning, structure at the conceptual level for recording and exchanging information on research inputs, activities, outputs, outcomes and impact indicators (http://casrai.org/program). The objective is to separate the labels, definitions, relationships and logical structures in our data from the specific local implementations of systems. Two key aspects of the CASRAI approach are (a) the ability, where needed, to extend the common terminology by discipline and by nation, and (b) the 'technology-agnostic' design that can support more than one format for local data storage and management. If the community can standardize at the conceptual level and integrate other standards and best practices (CERIF-relational, CERIF-XML, CRX, VIVO-rdf) at the physical database levels then a common approach is made more feasible and sustainable because we've removed the requirement of centralizing on a single central platform or database.

In addition is a need for creation and use of global standards to uniquely identify documents and people in the research enterprise: it is not enough to state that an individual is represented by a series of text-based data fields (Last Name, First Name, Middle Name). Multiple individuals can share a name, and individuals can change names. Publications are presented in different data sets with different formats and metadata. CrossRef is working to address this from the perspective of publications by creating DOIs, and more recently they have launched the FundRef initiative to standardize the acknowledgement of grant funding. Another player in this space is ORCID (http://orcid.org) a non-profit organization with an international and inter-sector membership and executive board which endeavors to create an identity system for scholars and researchers that supports the claiming of research "products" across data types and systems.

The goal of ORCID is to provide an international and interdisciplinary registry service for researchers and scholars and thereby support a permanent, clear and unambiguous record of scholarly communication. The ORCID identifier enables reliable attribution of authors and contributors by serving as a metadata tag on grants, publications, and other scholarly documents. Previous identifier schemes have failed because they were not tied into critical research workflows. OR-

CID has published Web services that may be used by academic institutions, funding organizations, publishers, and third-party vendors to integrate identifiers into processes such as manuscript submission, grant application, and employee profiles.

4 Security: Protecting data and providing access

A data infrastructure that has at its heart collection of data about individuals must address data privacy. Much of the data proposed for a workforce analysis data infrastructure are publicly viewable. However, some of these data when matched to a person record and linked to other data may have the potential to become sensitive. Other data may be sensitive from the outset, such as personally-identifiable information that may be included in some survey or administrative sources. Finally, there are data that may be protected by intellectual property or be otherwise considered proprietary. Privacy concerns can prevent stakeholders—individuals, governments, academia, and companies—from sharing their data.

Several models are in use to manage access to sensitive data. An example of a platform with multiple data sources, no user fee, and no access limitations are the The Interuniversity Consortium for Political and Social Research (ICPSR) or the Dataverse Network. Also with no user fee, but with specific licensing and security terms and conditions is the US National Center for Education Statistics and the US National Science Foundation. Research Data Centers, such as those provided by the US Census Bureau, provide restricted data licensed to cleared researchers to use at select locations and include a user fee. These centers provide excellent data security, but limit access to the research community due to the restricted geographic access and the expense of maintaining these centers (National Research Council 2005).

Another approach is to host public and restricted data using a multi-platform Web-based infrastructure and implement a hybrid access model. Only those users who have applied for an obtained security access to data sets would have access to them, but at the same time those data sets not covered by security restrictions would be available to users without licensing. An example of this is the forthcoming National Center for Science and Engineering Statistics' secure data access facility. An example of a repository managed by a private company is the Thomson Reuters MarketScan database, which contains information on medical claims data from public and private healthcare providers and payers, including Medicare, to support analysis of healthcare cost, treatment and the impact of patient behavior.

5 Conclusions

Stakeholders are driven by different definitions and reporting needs. Some require the ability to search for and qualify experts. Others may require ongoing updates of research activity to maintain internal personnel databases. Some organizations are seeking ways to reduce the burden of data entry and to concurrently improve data quality by providing services to ingest research activity data. Researchers interested in knowledge flow seek to study the linkages between people to better understand knowledge production, use, and impact. The recent Beyond Impact meeting (http://beyond-impact.org/?page_id=64) defined it thusly: the measure of how research outputs and outcomes influence and are re-used by other researchers, commercial partners, and wider society. Much of the work on research impact has been based on one output: research publications. Largely from the foundational work of Eugene Garfield (Garfield 1978), a number of variables derived from publications have proven to be relevant indicators of research activity. A data infrastructure would support development and testing of a broader array of impact indicators that cover not only more fields of endeavor but also more types of research outputs, with the ultimate goal of providing the best evidence base on which to support high quality decision making.

References

Garfield EG, Malin MV, and Small H (1978). Citation Data as Science Indicators. In *Toward a Metric of Science: The Advent of Science Indicators*, Eds. Yehuda Elkana, Joshua Lederberg, Robert K. Merton, Arnold Thackray and Harriet Zuckerman. New York: John Wiley & Sons.

Lai R, D'Amour A, Yu A, and Fleming L (2010): Disambiguation and co-authorship networks of the US Patent Inventor database. Available at: rd-dashboard.nitrd.gov/assets/disambiguation_of_uspto.doc.

Lane J (2010): Let's make science metrics more scientific. *Nature* 464:488-489.

National Research Council, *Expanding Access to Research Data: Reconciling Risks and Opportunities* (National Academy Press, Washington, DC, 2005). http://www.nap.edu/catalog.php?record_id=11434.

OECD (2007): Revised field of science and technology (FOS) classification on the Frascati Manual. DSTI/EAS/STP/NESTi(2006)19/FINAL. Available at http://www.oecd.org/dataoecd/36/44/38235147.pdf.

UNESCO (1976) SPINES: An international sstem for the exchange of inofmration on science and technology for policy-making, management, and development. UNESCO/NS/ROU/364/prev.Available at http://unesdoc.unesco.org/images/0001/000189/018980EB.pdf.

Wright, K, Williams, D, Schnell, J, and Haak, LL (2011): Linking grant portfolio investments to public health outomes using patent data. USPTO Patent Statistics for Decision Makers Conference, Alexandria, VA, USA November 15, 2011. http://conferences.thehillgroup.com/USPTO/PatentStatistics2011/index.html.

Contact Information

Laurel L. Haak, PhD (corresponding author)
ORCID, 10411 Motor City Drive, Suite 750, Bethesda MD 20817 USA
l.haak@orcid.org

David Baker
CASRAI, 200-440 Laurier Ave. West - Ottawa ON Canada K1R 7X6
dbaker@casrai.org

Matt A. Probus
Custom Analytics Division, IP and Science, Thomson Reuters , 1455 Research Blvd., 2nd Floor, Rockville, MD 20850 USA
Matt.Probus@thomsonreuters.com