The Web of Innovation: Using Website Data to Understand How Firms Innovate

Sanjay	Arora	sarora@air.org	USA	American Institutes for Research		✓	✓
Evgeny	Klochikhin	eklochikhin@air.org	USA	American Institutes for Research			

Firms are the core of the innovation system (Metcalf & Ramlogan, 2008). To produce innovation, they interact with other firms, individuals, universities, government, and other stakeholders (Etzkowitz, 2000). They also operate in a dynamic policy environment. While it is possible to recognize and study major determinants of firm innovation, such as networks and institutions, analyses often lack a dynamic perspective of unraveling the full innovation process leading up to market success or failure. As such, it is challenging to simultaneously assess specific aspects of the innovation journey including idea generation, research and development, and market-facing products and services (c.f., Rip, 2012).

How firms approach this innovation journey is a function of a diverse array of factors, including technology specialization and know-how, market receptivity, desire and ability to course-correct on strategic and tactical considerations, and access to alliances and networks. Most extant work examines these determinants through the use of traditional data sources that allow only some of these constructs to be analyzed at a given time. For example, journal papers and patent records describe only certain aspects of the innovation process, such as the production of knowledge, formal collaborations, disclosures of original product features and designs, and a desire to protect and enhance market share (De Bellis, 2009; Neal, Smith, & McCormick, 2008). Yet, the innovation process, and in particular its materialization in system settings, is a complex phenomenon involving many actors, institutions, laws and rules, and market forces.

Website data offers new ways to analyze many aspects of innovation simultaneously. Along with the traditional determinants of knowledge production and formal collaborations, these data can shed more light onto how and why firms innovate: Websites contain data on human capital, product portfolios, investments, and alliances and networks where firms are embedded as part of the larger innovation system (Arora et al., 2016). Albeit rich and easily accessible, website data are hard to structure and transform into an analytical format (Arora et al., 2013). As such, analyses of websites run into two major difficulties:

- Varying website structure: There is not a one-size-fits-all solution allowing researchers to scrape and organize data from multiple websites in a systematic way
- Challenging information extraction and variable operationalization: Unsophisticated methodological approaches may introduce significant systematic error when extracting information from unstructured website data (Arora et al., 2015)

Computer scientists have produced several approaches to overcome these challenges. In terms of handling varying website structures for the purpose of information retrieval (Penman, Baldwin, & Martinez, 2009; Grasso, Furche, & Schallhart, 2013), methods relying on machine learning algorithms can identify patterns in website structure and retrieve information that most interests the user. Researchers can consequently retrieve data from target websites in a customizable manner. We use one implementation of these novel web scraping techniques– Apache Nutch – to retrieve relevant data from firm websites and feed those data into subsequent analyses.

The key challenge is structuring the retrieved data in the needed analytical form. Based on existing literature, we identify variables pertinent to several analytical constructs: knowledge production; contingency operation; alliances and networks; human capital; product and technology specialization; and physical capital. For example, human capital can be explored as pertinent educational *or* reputational credentials of founders and team members, as disclosed on websites; knowledge production and expertise can be reflected in inventions found in patents or ongoing research activities as disclosed in publications.

We focus on three industrial sectors – nanotechnology, green goods manufacturing, and synthetic biology. In the sampling design, we apply three already-published keyword search strategies (Arora et al., 2013; Shapira et al., 2015; and Shapria et al., 2017) to USPTO PatentsView, an open data platform for US patent applications and grants, to produce the sample frame and to subsequently identify firms with valid website presences. We take into account any bias in this approach through descriptive analyses of the resulting dataset.

After selecting the firms and identifying their websites, we test two approaches to retrieve relevant variables: 1) structuring data as we crawl; and 2) organizing data once it is retrieved from the website and stored in a database. Information retrieval, natural language processing, and named entity recognition methods are applied to retrieve only relevant data points and draw associations between those variables in the analysis phase. By examining textual data, we build frequency tables and covariance matrices, as well as directed graphs, to employ two types of statistical models: (1) clustering and classification models, and (2) random walks and Markov chains. The clustering and classification models distinguish between different types of firms in the dataset (across and within the three industries). The random walk and Mark chain class of models probabilistically traverse the different associations between innovation variables to show different types of innovation strategies and orientations, as revealed through the website data.

The results show that high-tech small firms are likely to maintain distinct innovation strategies and orientations as revealed through their websites. For example, highly innovative small firms that patent frequently are more likely than non-highly patenting firms to stress certain concepts on their sites, e.g., university linkages, which signifies access to new knowledge and research and development capabilities. The results also show that firms with fewer ostensibly identified physical capital assets are less likely to be manufacturing firms and exhibit increased flexibility in terms of value capture approaches. Flexibility here was measured by way of website language conveying customization of configuration or products and/or services.