European Economic Review 36 (1992) 483–994. North-Holland

What is a firm?
A historical perspective

Alfred D. Chandler, Jr.

Harvard Business School, Cambridge, MA 02138, USA

1. Introduction

'What is a firm?' the title to this session asks. The answer seems to be easy. A firm is a legal entity – one that signs contracts with its suppliers, distributors, employees and often customers. It is also an administrative entity, for if there is a division of labor within the firm, or it carries out more than a single activity, a team of managers is needed to coordinate and monitor these different activities. Once established, a firm becomes a pool of learned skills, physical facilities and liquid capital. Finally, 'for profit' firms have been and still are the instruments in capitalist economies for the production and distribution of current goods and services and for the planning and allocation for future production and distribution.

I think most economists would agree on these four attributes of the firm. I, as a historian who has spent a career in examing the operations and practices of business firms, have had little trouble in locating information on literally hundreds of individual enterprises, each of which played a part in the creation and development of modern industries and modern economies. Nor do individuals have difficulty in defining the firms in which they work or the securities of those in which they invest.

If the firm is so easy to identify, why then do we have a session on 'What is a firm?' It is one that concerns economists more than economic historians, for it is less a question of economic practice than economic theory. Ronald Coase first raises the question years ago when he asked: If accepted theory assumes that the coordination of the flow of goods and services is done through the price mechanisms, 'why is such an organization necessary?' Therefore, he continued, 'our task is to discover why a firm emerges at all in a specialized exchange economy' (1937, pp. 388, 390).

As an economic historian I've concentrated on practice rather than theory. In *The Visible Hand* I investigated the beginnings and subsequent develop-

0014–2921/92/$05.00 © 1992—Elsevier Science Publishers B.V. All rights reserved

ment of what I term the modern multi-unit enterprise (a firm consisting of more than a single plant, shop, or office) in American transportation, communication, production and distribution. In *Scale and Scope*, published in 1990, I focused on the history of the modern industrial firm – the most complex and the most transforming of modern business enterprises – from the 1880s, when such firms first appeared, through World War II. I did so by comparing the fortunes of more than 200 of the largest firms in the three major indutrial economies – those of the United States, Britain and Germany – which until the Great Depression produced two-thirds of the world's output of industrial goods.

What I plan to do in this paper is: first to describe the similarities in the historical beginnings and continuing evolution of these enterprises, and then outline my explanation for these similarities. Finally, I relate my explanation of these 'empirical regularities' to four economic theories of the firm – the neoclassical, the principal–agent, the transaction cost, and the evolutionary. That is, I attempt to indicate the value of these theories for explaining the beginnings and growth of modern industrial enterprises.

2. Regularities described

The basic similarities in the collective history of the approximately 800 industrial firms were that a new type of enterprise appeared suddenly in the last two decades of the 19th century, that such firms continued to cluster in industries with the same characteristics throughout the 20th century, and that they were created and continued to grow in much the same manner. These industrial firms first appeared as modern transportation and communication networks were completed – networks that themselves were built, operated, enlarged and coordinated by large hierarchical firms. By the 1880s the new railroad, telegraph, steamship and cable systems made possible a totally unprecedented high volume, steady and regularly scheduled flow of goods and information through national and the international economies. Never before could manufacturers order large amounts of supplies and expect their delivery within, say, a week; or could they promise their customers comparable large-scale deliveries on some specific date. The new potential for greatly increased speed and volume of production of goods generated a wave of technological innovations that swept through Western Europe and the United States during the last decades of the 19th century creating what historians have properly termed the Second Industrial Revolution.

New industries appeared. Old ones were transformed. The making of steel, copper and aluminum; the refining of oil and sugar; the processing of grain and other agricultural products; and the canning and bottling of the products thus processed were all transformed. In chemicals new processes

produced man-made dyes, medicines, fibers and fertilizers. New mass produced office, agricultural and sewing machines quickly came on the market as did heavy machinery for a wide variety of industrial uses. The most revolutionary of the new technologies were those that generated and transmitted electricity for lighting, urban traction and industrial power. These new industries drove economic growth and played a critical role in the rapid reshaping of commercial, agrarian, and rural economies into modern, urban industrial ones. The newly formed enterprises that created and expanded these industries almost immediately began to compete in international markets.

Firms in these industries differed from older ones such as textiles, apparel, furniture, lumber, leather, publishing and printing, ship building and mining. They were far more capital-intensive, that is the ratio of capital to labor per unit of output was much greater. And they were able to exploit far more effectively the economies of scale and scope. In the new capital-intensive industries large plants had significant cost advantage over smaller ones. Up to a minimum efficient scale (determined by the nature of the technology and the size of the market) long-run cost per unit of output dropped much more quickly as the volume of output increased than was the case in the older labor-intensive industries. Many too benefited from the economies of scope – that is, those economies resulting from making a number of different products in a single factory or works using much the same raw and semi-finished materials and much the same intermediary processes of production.

Nevertheless, as I wrote in *Scale and Scope*:

> These potential cost advantages could not be fully realized unless a constant flow of materials through the plant or factory was maintained to assure effective capacity utilization. If the realized volume of flow fell below capacity, the actual costs per unit rose rapidly. They did so because fixed costs remained much higher and 'sunk costs' (the original capital investment) were also much higher than in the more labor-intensive industries. Thus, the two decisive figures in determining costs and profits were (and still are) rated capacity and throughput, or the amount actually processed within a specified time period.... In the capital-intensive industries the throughput needed to maintain minimum efficient scale required careful coordination not only of the flow through the processes of production but also the flow of inputs from suppliers and the flow of outputs through intermediaries to final users.
>
> Such coordination did not, indeed could not, happen automatically. It demanded the constant attention of a managerial team or hierarchy. The potential economies of scale and scope, as measured by rated capacity, are the physical characteristics of the production facilities. The actual economies of scale and scope, as measured by throughput, are organiza-

> tional. Such economies depend on knowledge, skill, experience, and teamwork – on the organized human capabilities essential to exploit the potential of technological processes (p. 24).

These enterprises in the new capital-intensive industries began and continued to grow in similar ways. All exploited the cost advantage of scale and scope. Nevertheless, investment in production facilities large enough to exploit these advantages were in themselves not enough. Two other sets of investments had to be made. The entrepreneurs organizing these enterprises had to create a national and then international marketing and distributing organization. They also had to recruit teams of lower and middle managers to coordinate the flow of products through the processes of production and distribution and teams of top managers to monitor current operations and to plan and allocate resources for future ones. The first firms to make the three-pronged set of investments in manufacturing, marketing, and management essential to exploit fully the economies of scale and scope quickly dominated their industries. Most continued to do so for decades.

The tripartite investment gave the first to make it – that is, the first movers – powerful advantages. To benefit from comparable costs, challengers had to construct plants of comparable size and do so after the first movers had already begun to work out the 'bugs' in the new production processes. The challengers had to create distribution and selling organizations to capture markets where first movers were already established. They had to recruit management teams to compete with those already well down the learning curve in their specialized activities of production, distribution, and (in technologically advanced industries) research and development. Challengers did appear, but only a few.

The three-pronged investment led to the creation of the modern multi-unit industrial enterprise in those industries where the cost advantages of the economies of scale and scope were the greatest. So from their beginnings in the 1880s they concentrated in the capital-intensive industries. The structure of these industries became, after a short shakedown period, and remained oligopolistic.

In the new or transformed capital-intensive, oligopolistic industries price remained a significant competitive weapon. But these firms competed even more forcefully through functional and strategic efficiency; that is, by carrying out more capably processes of production and distribution, by improving both product and process through systematic research and development, by locating more suitable sources of supply, by providing more effective marketing services, by product differentiation (in branded packaged products primarily through advertising), and finally by moving more quickly into expanding markets and out of declining ones. The test of such competition was changing market share and in the new oligopolistic industries market share and profits changed constantly.

Such oligopolistic competition in these capital-intensive industries sharpened the product-specific capabilities of workers and managers. These capabilities plus retained earnings from profits of the new technologies became the basis for the continuing growth of these managerial enterprises. Firms did grow by combining with competitors (horizontal combination) or by moving backward to control materials and forward to control outlets (vertical integration); but they took these routes usually in response to specific historical situations.

For most, the long-term continuing strategy of growth was expansion into new markets – either into new geographical or product markets. The move into geographically distant areas was normally based on the competitive advantage of organizational capabilities developed from exploiting economies of scale. Moves into related product markets rested more on capabilities developed from the exploitation of the economies of scope. Such organizational capabilities honed by oligopolistic competition provided the dynamic for the continuing growth of such firms, of the industries which they dominated, and of the national economies in which they operated.

3. Explaining the regularities

Thus the key concept I use to explain the similarities in the beginnings and growth of modern industrial enterprises is that of organizational capabilities. These capabilities were created during the learning process involved in bringing a new or greatly improved technology on stream, in coming to know the requirements of markets for new or improved products, the availability and reliability of suppliers, the intricacies of recruiting and training managers and workers. These capabilities were the collective physical facilities and human skills as they were organized within the enterprise. They included the physical assets of each of the many operating units – the factories, the selling and other offices, and the research laboratories – and of more importance the functional and administrative skills of the employees in such units. But only if these skills were carefully coordinated and integrated could an enterprise achieve the economies of scale and scope that were needed to compete and to continue to grow.

Such managerial skills were based on learning carried on within the different levels of the hierarchy – the operating units, the functional departments, and, as the firm grew, the product and geographical divisions and, of course, the corporate offices. Such learning was a process of trial and error, feedback and evaluation. It is more organizational than individual. Even the skills of individuals depended on the organizational setting in which they were developed and used. If these company-specific and industry-specific capabilities continued to be enhanced by constant learning about products, processes, customers, suppliers and other workers and managers

within the firm, the enterprise was usually able to remain competitive and profitable. If not, its market position deteriorated.

The creation, maintenance and expansion of such capabilities permitted American and German firms in the two decades before World War I to drive quickly British firms out of international markets and even Britain's own domestic one in most of the capital-intensive industries of the Second Industrial Revolution. They made it possible for German enterprises to regain swiftly their position in world markets after a decade of war, defeat and inflation between 1914 and 1924, and to come back again in the 1950s after a far more devastating war.

So too organizational learning permitted Japanese firms, first to carry out a massive transfer of technology from the west to Japan. Then once their domestic market became large enough to permit the building of enterprises large enough to exploit fully the economies of scale and scope, they developed organizational capabilities necessary to provide competitive advantage in international markets. Finally the economies that followed the Soviet model by relying on central planning agencies – Gossnap and Gosplan in Soviet Russia – to coordinate current flows of goods through the processes of production and distribution and to allocate resources for future production and distribution prevented managers in units of production and distribution from learning how to coordinate effectively flows of goods from suppliers and to markets based on information and knowledge about current facilities, available supplies and market demand. The failure to develop such capabilities has been central to the distintegration of these centrally planned economies.

4. Organizational capabilities and the theory of the firm

How then do established theories of the firm – neoclassical, principal–agent, transaction cost and evolutionary theories – relate to this historical description of and explanation for the development of the modern industrial firm that has transformed industries and economies in the past century and a half. How do these theories contribute to an explanation of economic growth and transformation?

The first two – the neoclassical and principal–agent theory – contribute little in their present abstract formulations. The neoclassical theory views the firm as a legal entity with a production set (a set of feasible production plans) from which a manager, acting rationally with full information, chooses the set most likely to maximize profits or present value of the firm [Hart (1989, p. 1758)]. Principal–agent theory accepts the neoclassical firm as a production set but gives it a managerial hierarchy. The advocates of this theory concentrate on the abilities of the 'owners' to discipline the managers with whom they have contracted to choose and implement the production

plans, but who may manage the firm in their own interest rather than for that of the owners. The proponents of agency theory concern themselves with the owners' problems of coping with asymmetric information, measurement of performance, and incentives. Both theories see the firm as a legal entity that contracts with outsiders – suppliers, dealers, financial institutions and the like – and insiders – workers and managers. But neither deals with the firm's physical facilities, and human skills and the resulting revenues on which the current profitability and future health of the enterprise depend.

Transaction cost theory has more relevance to the historical story and the explanatory concept of organizational capabilities precisely because it does incorporate investment in facilities and skills. Because it does, I have learned much from its practitioners, particularly Oliver Williamson. The theory focuses on transactions. As Williamson emphasizes, micro-economic activity is organized to economize on costs of production *and* transactions. At issue is whether the costs of transactions carried out by the enterprise are lower by relying on the market (where they are defined through contractual agreements) or by internalizing them within the firm. Such costs are reduced through internalization when firms make investments in highly specialized physical facilities and in human skills based on specialized learning. That is, these specialized assets can only be used for the production and distribution of specific products or services. So they lose value if deployed to other activities. This is particularly the case when a contractual arrangement involves many continuing transactions.

The reason is, Williamson argues, that long-term contracts are difficult to define, because the contracting parties cannot obtain all the necessary information. They act rationally, but this rationality is bounded. Moreover, the different parties involved, acting for their on self-interest, may suppress information, that is, they may act opportunistically often with guile. In Williamson's words: 'Any attempt to deal seriously with the study of economic organization must come to terms with the combined ramifications of bounded rationality and opportunism in conjunction with asset-specificity.' (1985, p. 42, also p. 30, 53).

The basic difference between myself and Williamson is that for Williamson: 'The transaction is the basic unit of analysis.' (1985, p. 41). For me it is the firm and its physical and human assets. If the firm is the unit of analysis, instead of the transaction, asset specificity still remains significant; but the specific nature of the facilities and skills are more significant than bounded rationality and opportunism to the shaping of decisions as to internalizing transactions and, therefore, in determining the boundaries between firm and market. For example, in the new capital-intensive industries the need for the firm to monitor high level throughput was much greater than in the older labor-intensive ones. Therefore, whereas firms in the capital-intensive industries internalized distribution, those in the labor-intensive ones continued to

rely on independent distributors [*Scale and Scope* (pp. 142–143, 153)]. Moreover, in capital-intensive industries the pressure to internalize varied with the source of supplies, nature of technology of production, and the size and requirements of markets. So too the pressure for backward integration varied as an industry grew and its leaders expanded into more distant markets.

An understanding of the specific characteristics of a firm's assets, particularly its learned organizational skills, is even more useful than an understanding of the impact of bounded rationality and opportunism on transactions involving those assets in explaining the continued growth of firms into new foreign and related product markets. Knowledge gained in creating a wholesaling organization led to the building of a comparable one in a foreign market. When such markets grew to a size that permitted the establishment of plants of minimum efficient scale, firms used their learned skills to build the new facilities. The number and location of the plants built abroad reflected the minimum efficient scale (mes) of the technology of production and the size of the markets. Thus steel, copper and aluminum plants where mes was very high were rarely built abroad; but in food and machinery many processing or assembling plants were built or acquired in foreign lands.

So too the organizational skills developed in one function often gave the firm a competitive advantage in a related product market. The move into new markets based on competitive advantage in one function required the building of complementary facilities and skills, and, in turn, trained managers in the ways of seeking out and capturing market opportunities. For example, in the years since World War II such organizational skills permitted business machine companies to become first-movers in mainframe computers, over-the-counter drug companies to become first-movers in the new antibiotic prescription drugs, chemical companies to move out of commodities into specialty chemicals and oil companies to replace chemical firms in the production of petrochemicals. Thus in analyzing the continued development of existing industries and the building of new ones, the firm would seem to be a more promising unit of analysis than the transaction, and the concept of the organizational capabilities that permit it to remain competitive, and therefore profitable, in national and international markets more pertinent than those of bounded rationality and opportunism.

This is why I'm sympathetic to the recently articulated evolutionary theory of the firm, first made explicit by Richard Nelson and Sidney Winter in their *An Evolutionary Theory of Economic Change* published in 1982. As Winter notes, their emphasis is placed on production rather than exchange. On the other hand, he continues, '[neoclassical] orthodoxy and transactions costs economics, place deal-structuring at center stage, and cast the economics of production and cost in a supporting role', (1988, p. 173).

The central concept of Nelson and Winter is that of routines. 'In

evolutionary economics, the specifics of the ways firms relate to owners, customers, and input suppliers are subsumed under the heading of organizational routines,' (1988, p. 173). They define '"routine" in a highly flexible way, much as "program" (or, indeed, "routine") is used in computer programming' (1982, p. 97). For them 'routines are the skills of the organization' that in turn become its 'genes' (1982, p. 134).

In a recent, still to be published, paper Nelson, building on his and Winter's past work and the more recent writings of David Teece, Giovanni Dosi, William Lazonick and myself, presents 'An Emerging Theory of Dynamic Firm Capabilities'. Here he focuses on 'three different, if strongly related features of a firm that must be recognized if one is to describe it adequately: its strategy, its structure, and its core capabilities'. For Nelson the strategy and structure of firms, not the transactions in which they are involved, shape its capabilities. For him strategy is 'what scholars of management mean, as contrasted with game theorists. I mean a set of broad commitments made by a firm that define and rationalize its objectives and how it intends to pursue them'. 'My concept of structure', he continues, 'also is orthodox, as is my belief that strategy tends to define a desired firm structured in a general way, but not the details. Structure involves how a firm is organized and governed, how decisions are actually made and carried out, and thus largely determines what it actually does, given the broad strategy'. Finally, '[s]trategy and structure call forth and mold organizational capabilities, but what an organization can do well has something of a life of its own' (1991, pp. 19–21).

5. Towards a dynamic theory of the firm

The emerging theory of dynamic firm capabilities is of great value to the economic historian, for it recognizes the centrality of the processes of production and distribution and of organizational learning in the creation, development and transformation of those processes. It also emphasizes differences in production and distribution technologies and activities of different industries and different sectors. For me personally this emerging theory will be of particular value in further historical analyses. By focusing more sharply on organizational learning, I should be able to say more about why functional and strategic competition in modern capitalistic economies play a larger role in changing market share and profit than does price, to explain more carefully the success and failure of enterprises to grow by moving into new regional or related product markets, and to analyze more precisely the competitive success or failure of national industries and even national economies in which these firms operate.

Like the builders of the evolutionary theory of the firm, I see agency and transaction cost theory of value to the economic historian but within the

framework of evolutionary theory. Like them, I am convinced that the unit of analysis in developing a relevant theory of the firm must be the firm, not the contractual arrangements or transactions that it carries out. Only by focusing on the firm can micro-economic theory explain why this legal, contracting, transacting entity has been in the past the instrument in capitalist economies for carrying out the processes of production and distribution, of increasing (or hampering) productivity, economic growth and transformation. Only by focusing on the firm can theory predict the firm's continuing role as an instrument of economic growth and transformation, and be of value for developing policies and procedures for maintaining industrial productivity and competitiveness in an increasingly global economy.

References

Chandler, A.D., 1977, The visible hand: The managerial revolution in American business (Belknap/Harvard, Cambridge, MA).

Coase, R., 1937, Nature of the firm, Economica 4.

Hart, O., 1987, An economist's perspective on the theory of the firm, Columbia Law Review 89, 1757–1774.

Nelson, R.R., Why firms differ, and how does it matter? Strategic Management Journal, forthcoming.

Nelson, R.R. and S.G. Winter, 1982, An evolutionary theory of economic change (Harvard, Cambridge, MA).

Williamson, O.E., 1985, The economic institutions of capitalism: Firms, markets and relational contracting (Free Press, New York).

Winter, S., 1988, On Coase, competence, and the corporation, Journal of Law, Economics and Organization 4, 163–180, Spring.