Berlin, le 25 janvier 2007

Evaluation et politiques: Y a-t-il de bons indicateurs pour la recherche?

(Introductory lecture at the WZB Conference, June 1-3, 2006
“Quality assurance of the sciences in transition”)

Robert Salais (Wissenschaftskolleg zu Berlin -WIKO, Berlin, 2005-2006)

“The audit society is a symptom of the times, coincidentally a fin de Siècle*, in which a gulf has opened up between poorly rewarded “doing” and highly regarded “observing””* (Michael Power, 1997, *The Audit Society. Rituals of Verification*, Oxford University Press)

Abstract. Evaluation through benchmarking processes is expanding quickly in firms as well as in national administrations, international organisations (or the construction of Europe). They cover a wide spectrum of domains, research as well economic, social or employment issues. Their basic tool is the use of indicators of performance (such as the global rate of employment to evaluate the performance of national employment policies). They draw their political legitimacy and acceptance from the fact that these are figures (hence a priori objective and non debatable). Issues are, in reality, much more complicated. Using indicators as a tool for governance is not only substituting technique for politics in the sake of political neutrality; it is at the same, even if often inadvertently, making politics through technical choice. By using some key European exemples, I put in contrast two conceptions for the building and use of indicators, an instrumental one (derived from New Public Management) and an ethical one (derived from the works of Amartya Sen). Then I try to apply these conceptions towards research evaluation and to sketch a pluralist approach to evaluation in terms of objectives, actors and methods.

Le propos de cette introduction n’est donc pas de contester l’utilité collective d’évaluer (y compris de manière quantifiée) la qualité et la pertinence des productions scientifiques. L’action publique a besoin de repères, de références sur lesquels

s'appuyer pour orienter son cours, pour mieux poursuivre et réviser ses objectifs, pour faire entrer les acteurs dans un processus d'apprentissage. Mon introduction a pour objet de montrer que c'est précisément au regard de ces besoins que l'emploi des indicateurs pose problème. Le problème présente trois aspects : élaboration des indicateurs, production des données nécessaires et usage dans la décision. Bien entendu, si le chiffre possédait sa valeur de vérité du fait de sa seule existence, s'il était par essence le reflet exact de la réalité qu'il prétend mesurer, il n'y aurait aucun problème. S'il en était ainsi, nous n'aurions collectivement qu'à suivre la direction qu'il nous donne. C'est, bien sûr, le rêve de la gouvernance optimale, celui de pouvoir passer d'un contrôle central à un contrôle indirect via des technologies de pilotage de la coordination économique et sociale, et ce à moindre coût. Vous qui, comme moi, fréquentez parfois le milieu de la décision politique, vous savez quel soulagement ce serait enfin pour les décideurs politiques de pouvoir affirmer (sans être contredit par les faits) que 'l'option politique A est meilleure que l'option politique B parce que sa performance est plus grande que celle de B'. Nous devons, en tant que chercheurs, résister à cette croyance positiviste, car nous savons qu'il n'en est rien, tout en sachant que son attraction est quasiment irrésistible au sein des décideurs politiques (*policy makers*). Comme j'espère avoir le temps de vous le montrer, la croyance positiviste conduit à une approche instrumentale des indicateurs. Ce qui risque de prendre le dessus dans la décision et le management politiques, c'est la recherche et la sélection de dispositifs d'action qui maximisent l'indicateur visé, indépendamment (et parfois au détriment) de l'amélioration réelle des situations.

I. Données de nature ou données sociales ?

Il faut d'entrée de jeu rappeler que les données statistiques (dont les indicateurs font partie) ne relèvent pas d'un état de nature. L'information à la base des décisions et de l'action n'est pas un donné ; elle est produite et sélectionnée. Les travaux fondateurs d'Alain Desrosières (qui ont fait l'objet d'un colloque au Centre Marc Bloch le 18 mai dernier) ont amplement précisé ce point. Permettez-moi de vous donner une référence essentielle : Alain Desrosières, *Die Politik der großen Zahlen. Eine Geschichte der statistischen Denkweise*, Berlin, Springer Verlag, 2005 (*La politique des grands nombres*, Paris, La Découverte, 1993).

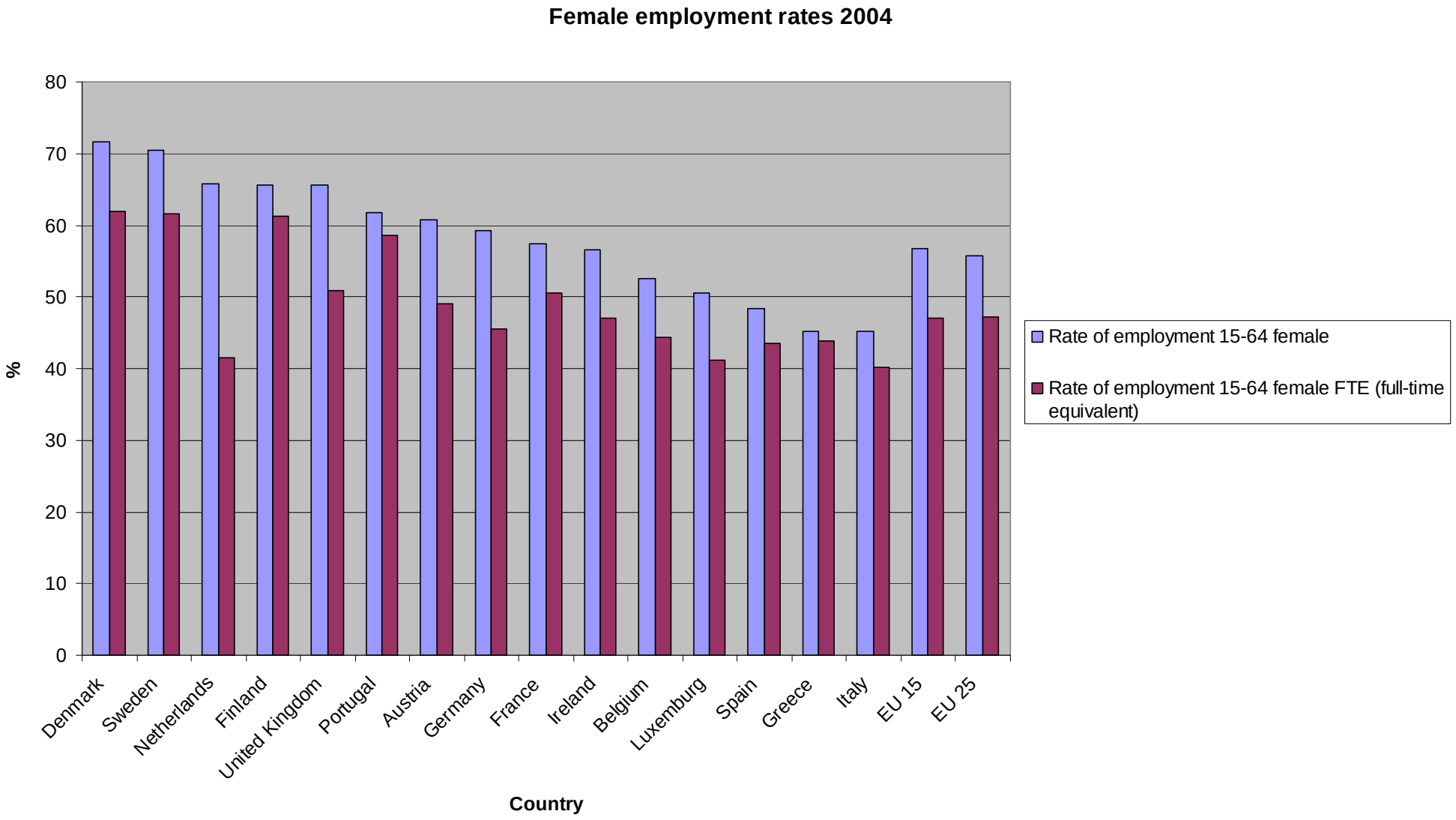
Lorsqu'on travaille, plus en ethnographe qu'en sociologue d'ailleurs, au sein de la machinerie qui les produit, on se rend compte que les données, précisément, ne sont pas « données » (donné = « what is taken for granted, or given »). Ce sont des produits, au bout d'une chaîne de production dont chacune des multiples étapes (ainsi que les relations qui les unissent) est sociale de bout en bout depuis la personne ou l'acteur objet de départ de l'observation jusqu'à l'agrégat ou le taux global. Les questionnaires et nomenclatures relatifs à un même fait social varient d'un pays à l'autre. Les statistiques administratives (mais aussi indirectement les enquêtes dans la mesure où les réglementations et droits sociaux influent les anticipations et les actions des intéressés, donc influent leurs façons de voir le monde et leurs réponses quand on les interroge) sont marquées par la législation sociale du domaine couvert (qui paie et selon quelles règles ; qui a droit à quoi et selon quelles conditions). Les catégories statistiques et administratives ont partie liée avec les catégories juridiques du domaine (et des domaines connexes). Les règles de gestion des agences ou des administrations formatent les données qui sont le sous-produit de leur activité. Ces règles sont hétérogènes d'un pays à l'autre, elles évoluent, elles sont manipulables, et manipulées parfois. Il s'ensuit, et c'est peut-être le plus important, que les catégories de perception des faits sociaux (comme le chômage) ou les critères de qualité (par exemple ce qu'est un « bon » travail scientifique ou encore les standards de qualité des produits dans l'industrie) sont historiquement et socialement inscrits (*embedded)*. Ces catégories et standards fondent les attentes (*expectations)* des membres d'une communauté à l'égard de la pratique sociale dans un domaine particulier ; ils varient dans le temps et d'un pays à l'autre. Il en est de même pour les modèles normatifs sur lesquels reposent les politiques publiques correspondantes. Toujours dans le domaine social, la norme est historiquement (et demeure en large part aujourd'hui) la responsabilité individuelle en Grande-Bretagne, celle de l'Etat en France, celle de la communauté de travail et de lieu en Allemagne, pour faire simple. A chiffres identiques, significations différentes.

Les enjeux d'une telle remise en perspective sont de deux ordres :

- L'ordre comparatif (Qu'est-il possible de conclure du fait arithmétique qu'un chiffre est plus grand ou plus petit qu'un autre ? Autrement dit sous quelles conditions deux données sont-elles commensurables ?)

- L’ordre normatif (Peut-on faire abstraction des normes qui sous-tendent une politique publique ? Autrement dit peut-on confondre mesure et évaluation ? Et si cette confusion est faite, quelles implications en découle-t-il pour la politique publique ?)

Il est essentiel de comprendre que ces deux ordres sont organiquement liés. Je voudrais tout de suite en donner un exemple, pris dans mon domaine de recherche.


Female employment rates 2004
80
70
60
50
40
30
20
10
0
%
Denmark
Sweden
Netherlands
Finland
United Kingdom
Portugal
Austria
Germany
France
Ireland
Belgium
Luxemburg
Spain
Greece
Italy
EU 15
EU 25
Country
Rate of employment 15-64 female
Rate of employment 15-64 female FTE (full-time equivalent)

Etre informé par la Commission européenne[1] que le taux d'emploi féminin en 2004 est de 65% en Grande-Bretagne et de 56% en France est une chose. Nous avons là deux chiffres dont l'un est, arithmétiquement, supérieur à l'autre, aucune personne sensée ne le contestera. En conclure que les performances en matière de participation à l'emploi des femmes sont meilleures en Grande-Bretagne qu'en France est une autre chose, nettement plus incertaine. Cette seconde affirmation n'a de sens que si tous les facteurs dont nous venons de parler et qui entrent dans la production de la donnée sont identiques, autrement dit, à supposer que ce soit possible tout du long, que les données pour être comparables soient recalculées à facteurs identiques. Par exemple, les pratiques et législation britanniques sont plus laxistes que les françaises à l'égard de ce qui doit être défini comme un emploi, par exemple quant aux conditions de travail à temps partiel (durée de travail et salaire spécialement). A l'issue d'une simple correction des différences de distribution des temps de travail, le taux d'emploi féminin, dit « équivalent temps plein », s'établit à 50% pour les deux pays ; les voici ex-aequo ! Et, d'une manière générale, comme le montre le graphique 1, le classement des pays européens auquel on arrive est foncièrement différent. Néanmoins, un satisfecit est délivré par la Commission Européenne au Royaume-Uni, « un des pays qui ont déjà dépassé l'objectif ». S'agissant de la France, il est noté qu'elle fait partie des « pays qui ont fixé des objectifs nationaux afin d'accroître la participation des femmes à l'emploi »[2]. Ce faisant, la Commission s'en tient donc à la statistique comme une vérité scientifique reflétant un état de nature qui ne peut souffrir discussion. Un emploi est un emploi ; l'homogénéité des situations d'emploi est totale, partout, en tout temps et pour tous les intéressés. Seule la performance quantitative ferait sens, appréciée de surcroît au niveau le plus global. Tout emploi, quel qu'il soit, est bon à prendre, car il accroît la performance.

Tout cela est bien surprenant, d'un point de vue scientifique ; en un mot on compare ce qui n'est pas comparable. Néanmoins, la Commission européenne et les organisations internationales persistent et même développent l'emploi des indicateurs. Une première réaction à cette bizarrerie est de considérer que nous en sommes encore à un stade infantile. L'orientation est bonne, mais il faut compléter la liste des

[1] Rapport conjoint sur l'emploi 2005.
[2] *Ibid.*, p. 53.

indicateurs et améliorer leur emploi. L'idée qu'il y a des bons et des mauvais indicateurs est affirmée, mais sur le strict plan de la démarche scientifique. On ne peut qu'approuver l'idée, mais certaines contraintes déterminantes issues de la pratique s'y opposent : spécialement l'exigence politique de base selon laquelle il faut des chiffres pour tous les pays et des indicateurs qui soient simples (pour être compris du public). Dans ce cas, l'obtention d'un « minimum commun » oblige à être peu regardant sur la production, la validité et le contrôle des données nationales. Lorsqu'il s'agit pour les administrations nationales d'expliquer à Bruxelles comment sont produites leurs données, il semble bien exister une sorte de limite invisible, mais respectée par tous, entre ce qui peut être porté à la connaissance de la Commission européenne et ce qu'elle n'a pas à connaître[3].

Sur le fond, je soutiendrai que la critique épistémique, nécessaire, ne va pas assez loin. Pour comprendre, il faut s'intéresser au lien entre évaluation et politique qui est mobilisé dans la pratique de ces organisations. Le fait est que celles-ci (et, au premier rang, la Commission européenne) agissent politiquement (et entendent agir) à travers des tableaux statistiques. La Commission européenne attend de la coordination des politiques nationales une amélioration de leurs performances quantitatives selon un ensemble d'indicateurs de *benchmarking* (et, indirectement, un reformatage de ces politiques qui optimise leurs performances). Elle a, pour ce faire, mis en place un énorme dispositif de procédures, de lignes directrices, d'objectifs décrits qualitativement et quantifiés par des tableaux d'indicateurs, tous terrains et tous azimuts. Ce dispositif de tableaux organise la coordination des politiques nationales en direction des objectifs. D'une certaine manière, c'est à cela que se réduit la coordination entre les Etats membres. Chaque année, il est demandé aux administrations nationales de remplir de données tous ces tableaux. C'est la matière avec laquelle, par exemple, la Commission évalue les Plans Nationaux pour l'Emploi (PNAE) ou la Stratégie de Lisbonne, les avancées, stagnations ou reculs et adresse ensuite à chaque Etat une série de recommandations (mélange de réprimandes et de propositions). C'est dans le jeu entre ce qui est demandé aux administrations nationales (et la manière dont c'est demandé) et ce qu'elles répondent (c'est-à-dire à

[3] Thedvall, R., 2006, *Eurocrats at work. Negotiating transparency in post-national employment policy*, Stockholm, Stockholm Studies in Social Anthropology, 58.

la fois ce qu'elles explicitent, mais aussi ce qu'elles taisent ou ignorent relativement au processus de fabrication des données des indicateurs) que s'insère la dimension politique du processus. A toutes les étapes du jeu, la politique passe par une multitude de choix techniques, le plus souvent petits, mais quelquefois fondamentaux (tel que le choix par le Centre de l'ensemble des indicateurs selon lesquels la performance va être mesurée). Le normatif – quelle conception de la politique publique ? – est omniprésent dans ces choix.

Les choix techniques faits ont des implications plus ou moins fortes. Un choix à forte implication est celui qui définit les conditions initiales, le cadre dans lequel la coordination se déroulera ; il contraint les trajectoires et les ajustements techniques futurs (qui auront, tous, un arrière-plan normatif). Un autre choix important est de mettre en avant le taux d'emploi et de disqualifier le taux de chômage comme cible privilégiée de la politique économique. On peut montrer qu'il lui correspond un basculement du modèle de plein emploi vers un modèle de dérégulation du marché du travail[4]. Un choix de portée plus limitée est, par exemple, de savoir si un demandeur d'emploi qui revient dans une agence, disons trois mois après en être sorti, doit être considéré comme un nouveau demandeur ou comme le même. Dans le premier cas, on améliore certains indicateurs de performance comme le taux de sortie hors du chômage ou la part des chômeurs de longue durée ; dans le second cas, on met en oeuvre une autre conception normative, celle de la qualité de l'emploi (en dessous de trois mois de durée minimum, une tâche de travail n'est pas un emploi). Les chiffres seront moins bons, mais l'exigence envers la politique du marché du travail plus élevée.

Vous comme moi, avons eu bien des fois à remplir des tableaux d'indicateurs de performances de nos laboratoires de recherche. Et à nous demander comment compter telle publication selon son standard, si l'auteur faisait bien partie ou pas du laboratoire, que faire des publications à multiples auteurs, des multiples variantes d'à peu près la même chose, etc. Nous avons aussi appris à enjoliver le résultat, sans le trahir pour autant. Que penser, dans ces circonstances, du chiffre agrégé sur la qualité

[4] Salais, R., 2006, « Reforming the European Social Model and the politics of indicators : from the unemployment rate to the employment rate in the European Employment Strategy », in Jepsen, M. and Serrano A., eds., 2006, *Unwrapping the European Social Model*, Bristol, The Policy Press, p. 189-212.

de l'effort national de recherche qui arrive sur la table d'un Ministre ou d'un Commissaire ? Que penser aussi des indicateurs et de leur sélection ? Il y a là, dans le domaine couvert par la Conférence, bien des recherches à faire. La priorité serait de pouvoir connaître dans le détail toutes les étapes de l'agrégation, depuis la production des données élémentaires par activité, poste de dépense, contenu de la dépense jusqu'à la valeur globale de l'indicateur. La comparabilité n'apparaîtrait plus aussi évidente.

II. Deux approches de l'évaluation : instrumental versus éthique

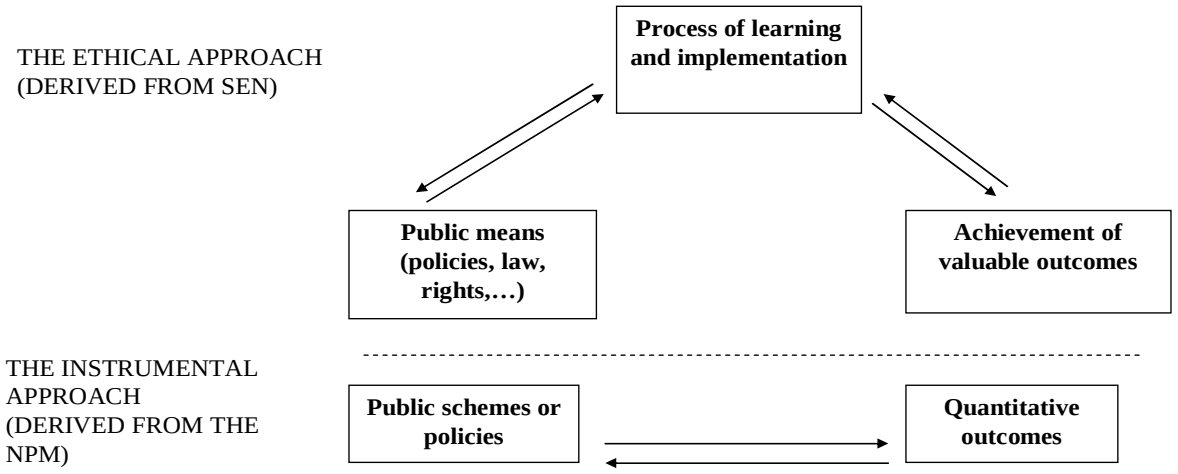
Juger d'un indicateur s'il est bon ou mauvais dépend de ce qu'on veut en faire, de la concurrence avec d'autres méthodes, du processus par lequel il est élaboré et choisi, du processus de fabrication des chiffres correspondants, des conclusions qu'on en tire et comment il rétroagit dans le processus de décision politique. Pour mettre un peu de clarté, cette partie oppose deux modes d'emploi (au sens large) des indicateurs dans un processus politique, l'un instrumental, l'autre éthique[5].

L'essentiel du propos s'appuie sur le Diagramme 1. Ce Diagramme oppose deux types de rapport entre politique publique et évaluation : un rapport causaliste et instrumental (2.1) ; un rapport médiat et éthique (2.2).

[5] More developments in Salais, R., 2006, « On the correct (and incorrect) use of indicators in public action », *Comparative labor law & Policy Journal*, vol. 27, 2, Winter, p. 237-256.

Diagram 1. Two conceptions of the relationship between public policy and evaluation

Improving the power of conversion of means into achievement of valuable outcomes



Maximising the statistical profile

2.1. La conception causaliste et instrumentale des indicateurs

La conception causaliste et instrumentale est représentée par la Nouvelle Gestion Publique[6] ou plutôt, car l'inspiration est définitivement anglo-saxonne, par l'école (certes diversifiée en son sein) du *New Public Management* (NPM). C'est la partie inférieure du Diagramme 1. La source est le management de l'entreprise (spécialement dans les techniques d'étalonnage, ou plutôt, encore une fois, de *benchmarking*). Ces méthodes poursuivent des visées essentiellement instrumentales. Ce qui importe pour ces méthodes, c'est le résultat, la performance. Les indicateurs sont utilisés pour mesurer la performance et pour classer, à des fins de mise en concurrence, les entreprises ou toute autre instance (organisation, politique, …). Toute réforme ou amélioration de l'existant est bonne quelque soit la nature des moyens utilisés (notamment leur fondement ou implication en terme de justice sociale, de normes internationales ou de droits fondamentaux) du moment qu'elle accroît la performance à coût constant. En termes plus théoriques, la NPM et les méthodes dérivées fonctionnent selon une approche conséquentialiste. Il est possible, selon elles, d'établir le schéma de causalité directe qui va des résultats quantitatifs atteints sur les indicateurs retenus aux moyens utilisés. Ce schéma supposé de causalité est utilisé, sans aucune médiation, pour ajuster les moyens lorsque la performance sur les indicateurs de gestion est mauvaise. Il s'agit donc moins d'<u>évaluer</u> (au sens de prendre comme référence un ensemble de valeurs) que de <u>mesurer</u> et de chercher les variables d'action qui permettraient d'améliorer la performance. Dans les théories du management de la firme, la chose se présente de la manière suivante[7]. Dans un contexte concurrentiel en évolution, une firme dépassée doit, pour survivre, faire évoluer son organisation via l'imitation des routines innovantes de ses concurrents (les « bonnes pratiques »). Elle doit identifier de manière correcte quelles sont ces routines innovantes, les transférer ensuite adéquatement dans sa propre organisation, arriver enfin à mobiliser son personnel pour s'approprier ces routines et les mettre en œuvre efficacement. L'ossature de la MOC européenne en est proche et s'en inspire (voir, par exemple, le vocabulaire des lignes directrices et des plans – une approche bottom-up, mais initiée et pilotée top-down).

---

[6] Excellente présentation dans Frédéric Varone et Jean-Michel Bonvin (s. dir.), La nouvelle gestion publique, numéro 1-2 de *Les politiques sociales*, 2004.

[7] L. Tronti, L., « Fruitful or Fashionable ? Can Benchmarking Improve the Employment Performance of National Labour Markets ?, in Best, E. et Bossaert, D., *From Luxemburg to Lisbon and beyond. Making the Employment Strategy Work*, Maastricht, European Institute of Public Administration, 2001, p. 67-83.

La difficulté des problèmes à résoudre (identification des bonnes pratiques, transfert, création d'un consensus) apparaît lorsqu'on réfléchit à ce qu'implique la transposition de cette méthode de management de la firme aux administrations publiques et, au-delà, à la coordination des politiques nationales par un Centre qui, à la différence de la direction d'une entreprise, n'a aucun pouvoir hiérarchique pour décider et impulser. La résolution du premier – identifier les bonnes pratiques- conditionne celle des suivants. Il faut mesurer les écarts de performance entre firmes (ici, entre politiques nationales) d'une manière cohérente et viable. Ces écarts doivent être corrélés adéquatement aux objectifs de la firme (ou des politiques) et pointer sur les causes visées, i.e. les bonnes pratiques. C'est à ce point décisif qu'entrent en scène les indicateurs utilisés. Si, pour une firme, un accord peut être assez facilement obtenu sur les objectifs (profit, croissance, réduction des coûts, valeur de l'action, productivité, etc.) et leur quantification et certaines relations mises en évidence entre eux, la question est autrement ardue pour des politiques publiques. Leurs buts sont multiples et contradictoires. Un même problème peut être identifié et résolu de plusieurs manières selon le pays. L'écheveau des interactions qui conduit de la performance quantitative à l'identification des « bonnes pratiques » bien souvent ne peut être démêlé. Enfin, il ne faut pas oublier le lien du NPM avec le VFM auditing (Value for Money). L'équivalence cherchée entre situations est immédiatement générale (une quantité globale) et même monétaire (monnaie). Peut-on ainsi négliger tous les niveaux intermédiaires où établir et discuter de la mesure serait nécessaire ? De telles interrogations se transposent sans difficulté quant à l'évaluation publique de disciplines scientifiques qui ont des méthodologies et des critères de qualité, des traditions propres, ou qui ont un rapport spécifique à la quantification. Nous le discuterons partie III.

On peut peut-être se passer d'une délibération au sein de l'entreprise. En revanche, la délibération politique et sociale, la plus large et la plus approfondie possible, semble nécessaire pour obtenir un accord ou un compromis minimum sur les valeurs et les normes au principe des décisions publiques. Cela suppose, bien sûr, de garder l'idée que toute décision publique ne peut s'abstraire d'une mise en relation avec des objectifs fondamentaux. S'affranchir de ces contraintes (délibération et référence à des objectifs fondamentaux) conduit à nier la distance qualitative (et interprétative) qui existe entre catégorie européenne et catégories nationales, ou entre la dynamique de la firme par rapport à sa référence, ou entre les standards de qualité employés par les diverses disciplines scientifiques. Or dans ses stratégies, la Commission européenne a mis d'entrée de jeu hors du débat public les questions

essentielles : mesurer quoi ? Et par quelles méthodes ? Il se peut très bien – c'est l'inquiétude et même la critique faite par Lundvall et Tomlinson (2002)[8] à la Stratégie de Lisbonne – que les capacités des firmes les plus nécessaires à la Stratégie soient intraduisibles en terme d'indicateurs. L'apprentissage est alors détourné vers la découverte de moyens instrumentaux pour améliorer, non pas la performance réelle, mais directement le score en terme d'étalonnage statistique. A terme les effets en termes d'efficience risquent dans ce cas d'être désastreux, sans parler de ceux en termes de justice sociale.

2.2. La conception médiate et éthique des indicateurs

La seconde conception des indicateurs et du rapport à établir entre politique publique et évaluation accorde, au contraire, tout son poids au fait d'évaluer. L'évaluation implique une référence explicite à des valeurs. C'est un schéma ternaire, et non binaire[9] (partie supérieure du Diagramme 1).

Les résultats visés par les politiques nationales sont, dans cette conception, le degré de réalisation (autrement dit, de mise en œuvre effective) par chaque pays de principes et d'objectifs fondamentaux. Le concept de mise en œuvre effective est complexe, mais important. Ce qui doit être évalué vise précisément le degré auquel une norme (par exemple, un standard de qualité scientifique) est devenue une institution réelle, le degré auquel elle s'est incorporée dans les pratiques et les attentes économiques, politiques et sociales au sein du pays considéré.

Le point de départ pour l'emploi des indicateurs est aux antipodes de l'a priori d'homogénéité des situations nationales entre elles. Pour bien évaluer, il faut respecter les idiosyncrasies ; plus même l'appréciation doit être conduite par les acteurs « locaux » eux-mêmes, sous certaines conditions de délibération publique. Car, ainsi que le soulignent les travaux cités en introduction sur la statistique, il existe un gap qualitatif entre une catégorie générale – qu'implique tout indicateur – et des catégories locales, nécessairement ancrées dans des

[8] Lundvall, B. and M. Tomlinson, 2002, "International benchmarking as a policy learning tool" in M.J. Rodrigues, ed., 2002, *The New Knowledge Economy in Europe*, Cheltenham, Edward Elgar, p. 203-231.

[9] Cette seconde conception doit beaucoup à ma lecture des travaux d'Amartya Sen. Voir Amartya Sen, *Development as Freedom*, Oxford, Oxford University Press, 1999; Robert Salais et Robert Villeneuve, (s. dir.), *Europe and the Politics of Capabilities*, Cambridge University Press, Cambridge, 2004, p. 283-300.

processus historiques et sociaux propres à chaque pays (ou dans des trajectoires spécifiques à des communautés de recherche). Le savoir pratique sur une situation d'action est distribué. Il est construit, de plus, selon des points de vue et des intérêts multiples. L'exigence de combiner l'emploi de catégories générales avec la prise en compte des singularités pertinentes de la situation implique nécessairement une délibération des choix collectifs. Cette délibération doit disposer de procédures *locales* où soient acteurs ceux qui connaissent les situations (précisément de ceux qui en ont l'expérience, qui possèdent une part du savoir pratique). De telles procédures locales devraient viser à découvrir les acteurs porteurs de ces savoirs, à les mettre en position de mobiliser ces savoirs pour concrétiser les objectifs d'une manière qui soit adéquate à leur réalité. Accroître l'emploi – en respectant des standards de qualité – peut signifier, ici mettre en avant l'effort et la qualité de l'innovation, là la découverte de nouveaux marchés, ailleurs avoir des procédures efficaces de reconversion des entreprises, etc. Autrement dit le lien objectifs – moyens doit être à chaque fois reconfiguré, et ce selon une délibération autonome entre acteurs pertinents.

Il est vrai (pour prendre l'exemple de la recherche) que ni les intérêts, ni les finalités, ni les savoirs sont identiques entre les scientifiques, les industriels qui mettent en œuvre des dispositifs issus de la recherche appliquée, les usagers, les gestionnaires administratifs et financiers, les défenseurs de l'éthique ou de l'environnement, les responsables politiques, etc. Néanmoins on doit considérer que tous, selon des formes et des degés divers, comptent. La mobilisation de leurs connaissances permettrait d'apprécier la distance entre objectif général et réalisation effective dans leur domaine, de repérer les dimensions saillantes et de proposer des indicateurs adéquats du degré de réalisation des objectifs. Ces indicateurs –contrairement à ces tableaux que, nous chercheurs, devons aujourd'hui remplir – n'ont aucune raison d'être a priori semblables d'un lieu à l'autre, d'un domaine à l'autre.

Ce sont là les caractéristiques de ce que j'appellerai une action publique située[10]. Cette seconde conception prend une posture moins triomphaliste que la première. Elle reconnaît que les indicateurs de performance issus de l'approche causale ne peuvent mesurer ces degrés de réalisation, même s'ils ont, bien sûr, un rapport plus ou moins direct, plus ou moins biaisé avec eux. Elle ne nie pas leur utilité, mais à titre subordonné à des finalités plus essentielles.

[10] See Storper, M. and Salais, R., 1997, *Worlds of Production. The Action Frameworks of the Economy*, Cambridge MA, Harvard University Press, part III.

Les politiques publiques n'existent pas pour améliorer leurs résultats, conception qui met la référence de l'évaluation du côté de l'efficacité organisationnelle de l'Etat dans une boucle auto-référentielle. Leur objectif doit être recherché du côté d'objectifs fondamentaux : justice sociale, développement humain, équité, progrès de la liberté réelle (évaluée selon toute l'étendue des droits économiques, politiques et sociaux), progrès de la connaissance.

L'enjeu essentiel de l'évaluation est remonté vers la « zone » médiane, là où, pour le progrès des réalisations, doivent interagir la mise en œuvre des politiques, la découverte des réalités et la construction d'indicateurs adéquats (zone désignée dans le schéma par « processus d'apprentissage et de mise en œuvre »). Nous ne sommes plus dans une causalité mécaniste et externe, mais dans un processus humain collectif. Le progrès des résultats finaux en termes de réalisation ne peut être isolé du développement des capacités d'action et d'initiative de chacun des participants à ce processus. Du bon agencement interne de ce processus dépendront, d'une part l'évolution des politiques publiques nationales et l'affinement de leurs outils, d'autre part le progrès des réalisations et celui des connaissances sur les situations de mise en œuvre, enfin une mise en rapport efficace des réalisations et des politiques. Contrairement à la notion de « bonnes pratiques » de l'école NPM, il s'agit moins pour un pays (ou une discipline scientifique) d'importer de l'extérieur des routines que d'apprendre par soi-même (par réflexivité) à évoluer en toute autonomie vers plus de réalisation. Rien ne se fera, au-delà des apparences, sans l'initiative autonome des participants (et, au premier rang, dans nos domaines, celle des scientifiques). Que ce soit cette initiative qu'il faille déclencher doit être la prémisse de toute action démocratique attachée à l'universalité de ses objectifs.

Ma présentation demeure trop lacunaire pour prétendre recenser toutes les questions qui se posent. La distinction entre ces deux approches instrumentale et éthique est importante de mon point de vue. On constate par divers signes que l'approche instrumentale tend à se cacher derrière une rhétorique empruntée à l'approche éthique. Surtout, cette distinction permet d'analyser les dispositifs d'évaluation existants comme combinant, à des degrés divers, les deux modèles, intrumental et éthique. J'esquisserai dans la partie suivante les grandes lignes d'une telle analyse à propos des dispositifs d'évaluation de la recherche.

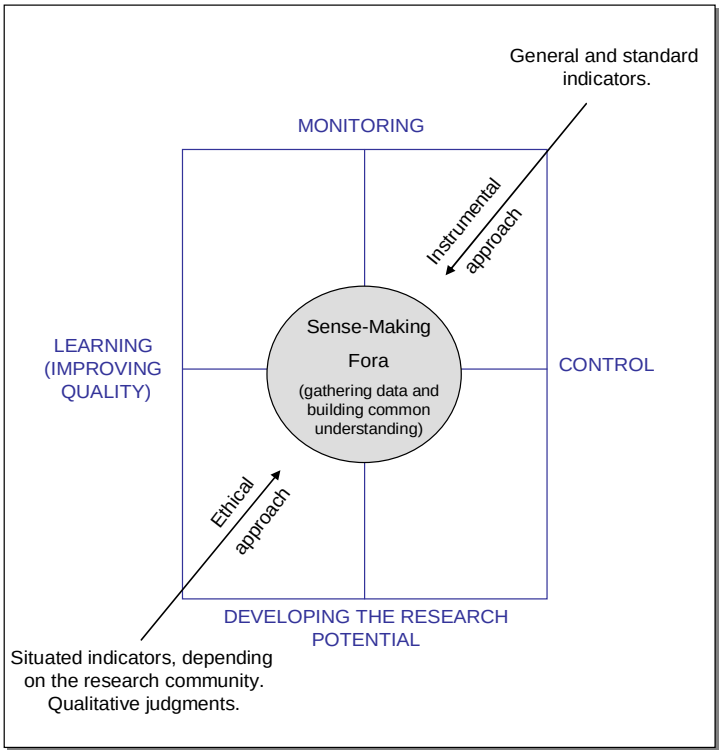
III. Une approche pluraliste de l'évaluation de la recherche (objectifs, acteurs, méthodes)

Une analyse réaliste de l'évaluation de la recherche scientifique doit prendre en compte la pluralité des objectifs à laquelle elle doit simultanément satisfaire ainsi que la diversité de ses acteurs. Partant d'une représentation simplifiée des objectifs et des acteurs (schéma 2), on tentera de mettre en œuvre, à titre d'essai, la distinction entre approches instrumentale et éthique élaborée précédemment et d'en tirer quelques conclusions provisoires.

Le schéma 2 distingue quatre objectifs de l'évaluation selon une configuration que nous allons expliquer : le contrôle, le pilotage (*monitoring*), l'apprentissage (*learning*), le développement[11]. Les procédures d'évaluation, habituellement, mettent en jeu plusieurs acteurs : les chercheurs[12] bien sûr, les financeurs, les administrateurs de la recherche et, bien qu'en général ils ne participent pas directement aux procédures, mais font entendre leur voix, les utilisateurs (citoyens, entreprises, collectivités publiques, etc.). On peut aussi retenir l'idée que tous ces acteurs sont d'accord sur la nécessité de poursuivre ces objectifs. Mais ils diffèrent sur leur hiérarchisation et, le plus souvent, accordent la priorité à l'un (parfois deux) d'entre eux sur les autres. Il s'ensuit que chacun a sa propre conception des autres objectifs, déduite de l'objectif qu'il considère comme prioritaire.

[11] Ce schéma est adapté de Engel, P , Carlssson, C. and van Zee, A., 2003, « Making evaluation count : Internalising evidence by learning », *Policy Management Brief*, European Centre for Development Policy Management, 16, August

[12] Définis ici, non sous l'angle de la profession de chercheur, mais comme tous ceux qui font des recherches, indépendamment de leur statut professionnel (professeur, chercheur, doctorant, post-doc, etc.), pourvu que leur activité soit labellisée comme recherche.



Scheme 2. Combining approaches and objectives of evaluation process for research activities

En première approximation, on peut considérer que les deux premiers objectifs (contrôle et monitoring, à droite et en haut du schéma) relèvent de manière dominante de l'approche instrumentale, même si d'autres préoccupations apparaîtraient à l'observation.

On peut à l'expérience, et sans grand risque de se tromper, attribuer aux financeurs comme objectif dominant celui du contrôle. L'argent a-t-il bien été employé selon les règles de transparence (dépenses et calendrier conformes au plan de charge) et d'attribution claire des responsabilités (qui fait quoi, quand, avec quel coût et comment) ? On reconnait là aisément, par exemple, la structuration des dossiers que les programmes de recherche de l'Union européenne demandent de remplir pour tout projet. Le second objectif (monitoring) est l'objectif dominant des administrateurs de la recherche. Ayant fixé des grands objectifs et

thèmes de recherche, répondant aux priorités politiques définies par leur gouvernement ou l'Union européenne, ils ont à cœur de suivre leur bonne mise en œuvre et leur achèvement.

S'inscrire dans l'approche instrumentale est naturel pour de telles fonctions qui visent un usage économe de la dépense publique. Les agences et autorités correspondantes ne se désintéressent pas des deux autres objectifs d'apprentissage (qui conditionne l'amélioration de la qualité de la recherche) et de développement du potentiel (national ou européen) de recherche. Mais elles les voient comme devant obéir à des critères conduisant à des choix rationnels qui peuvent être défendus avec des arguments objectifs dans une négociation budgétaire. La plupart des batteries d'indicateurs qui sont proposées par différentes instances nationales et internationales appliquent de ce fait une logique de performance et de classement. Les critères recherchés sont généraux et quantitatifs. Ils semblent offrir une sécurité du jugement sur des activités complexes et aux résultats incertains. Ils sont en fait manipulables.

Un grand nombre de ces indicateurs de performance sont en relation avec les publications ou citations dans des journaux scientifiques, avec des positions de referee dans les revues. La composition des listes de revues considérées comme de haut niveau, la prise en compte des ouvrages collectifs issus de colloques ou des manuels, le genre de citation, les débouchés vers des publics citoyens ou des acteurs économiques et sociaux sont autant de problèmes qui, selon la solution prise, conduisent à modifier la hiérarchie des performances, donc les suites données à la sélection des projets ou des chercheurs.

Les coûts pour la collectivité de ce type d'évaluation sont à la hauteur des avantages espérés de simplicité et d'objectivité. On retrouve des coûts de même nature que ceux vus précédemment dans la méthode ouverte de coordination européenne. En témoignent de multiples observations qu'on trouve au fil des rapports consacrés à ces questions[13]. L'apprentissage de la qualité tend à être détourné vers des comportements rationnels d'optimisation du score : en terme de nombre de publications par le biais de la duplication d'articles proches les uns des autres, mais adressés à des revues différentes (ainsi que

[13] Council for the Humanities Social Sciences Council, 2005, *Judging research on its merits*, Report, Amsterdam, Royal Netherlands Academy of Arts and Sciences; L. Hantrais, 2006, *Pour une meilleure evaluation de la recherché publique en sciences humaines et sociales*, Conseil National d'Evaluation de la Recherche (CNER), Paris, La Documentation française.

d'orientations vers certains types de publication au détriment d'autres) ; plus grave, en terme de choix des domaines et voies de recherche. Face à une sélection sévère et aux rapports de pouvoir au sein des disciplines ou entre elles, des jeunes chercheurs talentueux préfèrent choisir les voies bien balisées et au rendement sûr, plutôt que des voies innovantes, mais hasardeuses et mal repérées dans les procédures d'évaluation. Choisir l'interdisciplinarité cumule tous ces dangers. Le développement du potentiel national de recherche risque d'être orienté vers les domaines et les projets promettant les scores les plus élevés et de ne pas rencontrer les besoins réels de recherche de l'économie et de la société[14]. Ce danger serait encore plus grand si les procédures de financement de la recherche utilisaient des incitations fondées sur des indicateurs de performance. On risquerait de tourner en rond au sein d'une boucle autoréférentielle où chacun serait content, mais où aucun progrès réel ne se produirait. Les choix techniques d'indicateurs ont ainsi des conséquences politiques inattendues et non maîtrisées.

Les deux autres objectifs (l'apprentissage – autrement dit l'amélioration de la qualité de la recherche, et le développement du potentiel de recherche, à gauche et en bas) relèvent de manière dominante de l'approche qualifiée plus haut d'éthique. Car ce sont des objectifs fondamentaux de toute collectivité (nationale ou européenne considérée). L'évaluation devrait avoir pour finalité essentielle leur réalisation et leur internalisation dans l'activité de recherche. Il n'en reste pas moins, et c'est même primordial, qu'il faut porter un jugement sur la qualité (sans refuser, pour autant, l'aide de la mesure) et évaluer les potentialités de développement et l'adéquation de leurs orientations. Les critères à trouver ne peuvent être que des critères endogènes à la pratique individuelle et collective de la recherche. Ceci suppose de partir des réalités de la pratique.

En gros, la pratique est organisée en communautés[15] de recherche, par disciplines et, au sein de chaque discipline, par approches (diverses mais légitimes, que ce soit en termes d'objet, de problématique, de théorie ou de méthodologie). Ce sont des communautés de niveau et de taille nationale et, de plus en plus, internationale, dotées de leurs propres instruments de

[14] Les départements des Universités britanniques, par exemple, intègrent dans leurs critères de recrutement la capacité à être un « chercheur actif », en entendant par là l'habilité à obtenir des scores élevés dans les indicateurs nationaux d'évaluation du système universitaires. Certaines Universités organisent des formations à cette aptitude, pour les nouveaux recrutés.

publication et de référence. Il s'ensuit une conséquence importante : les références de jugement sur la qualité et, plus généralement, les critères d'évaluation sont endogènes à chaque communauté de recherche. Les références pour évaluer la qualité, en particulier, sont liées au processus de recherche propre à chaque communauté, à l'avancement de son front avancé de recherches qu'elle est seule à connaître intimement de l'intérieur. Il existe des recoupements et des chevauchements des programmes de recherche ainsi que des recouvrements des critères et des indicateurs. Mais il n'y a aucun raison de considérer que, du point de vue de l'apprentissage de la qualité, un critère ou une batterie de critères est a priori supérieur à une autre, ni qu'il faudrait imposer une batterie homogène de critères à toutes les disciplines et, en leur sein, à toutes les approches. Si indicateurs il doit y avoir, leur définition et leur choix doivent partir des propositions de ces communautés et suivre un processus délibératif, et non être imposées d'en haut.

On rencontre ici la question du jugement par les pairs, non pas comme alternative à la quantification, mais comme composante nécessaire de l'évaluation (la composition entre les deux pouvant varier selon la communauté de recherche). L'expression n'est pas tout à fait heureuse, car elle peut renvoyer à des effets corporatistes ou de clique visant à s'assurer le contrôle des moyens (et à se les partager discrètement). Néanmoins, dans une collectivité de recherche qui fonctionne et qui maintient sa déontologie, le jugement par les pairs est, pour un chercheur ou un projet, la sanction la plus dure et la plus rigoureuse, scientifiquement et dans ses effets[16]. Car cette collectivité est aussi un espace de compétition et d'innovation. Le jugement est le fait de personnes à qui « on ne la fait pas », tant ils connaissent de l'intérieur les travaux, les lignes de recherche, les réputations.

Une partie de l'expression et de la connaissance des besoins des utilisateurs (des acteurs économiques et sociaux, des décideurs publics) se fait auprès de ces collectivités de recherche. Les échanges, la négociation et l'exécution des contrats, la participation des acteurs aux protocoles de recherche ou des chercheurs à la formation des acteurs (par exemple dans

15 Le mot est un peu fort, car les frontières et les appartenances sont évolutives et n'excluent pas des recouvrements. Il est employé faute de mieux.

16 Les chercheurs continuent de préférer le jugement par les pairs. Et il est difficilement remplaçable par des indicateurs bibliométriques. L'INSERM en France a récemment comparé ces deux types d'évaluation pour 273 équipes de recherche. La corrélation statistique est faible, ce qui signifie que les pairs ne sont pas aussi influencés par l'output en terme de publications qu'on ne l'imagine. Voir N. Haeffner-Cavaillon, 2006,

l'entreprise ou la branche) servent aussi à cela. Par ce biais, les besoins sont traduits en un langage commun qui rend possible une recherche, à la fois de qualité et utile. La veille et l'anticipation des besoins doivent mobiliser, spécialement, les administrateurs de la recherche, mais ils ont aussi besoin de s'appuyer sur ces savoirs « locaux ». Enfin, on ne s'avance pas trop en notant que la pratique de la recherche induit une certaine conception du contrôle et du monitoring. Le contrôle souhaité est fondé sur la confiance et la souplesse. L'encadrement par les règles de gestion et de financement est accepté dans la limite où sa prise en compte améliore le protocole de recherche. De même les communautés de recherche entendent participer, sous une forme ou une autre (à travers leurs représentants ou leurs chercheurs les plus réputés) à la définition des grands objectifs et à leur suivi.

On peut conclure de ce rapide panorama que l'accord sur les objectifs et leur pluralité s'accompagne de la concurrence entre diverses conceptions des fonctions et des méthodes de l'évaluation. Les conflits d'intérêt sont donc la règle et le consensus (le vrai) l'exception. L'innovation, paradoxalement, est difficile, car elle doit trouver son chemin dans un univers très structuré par des règles, des intérêts, des communautés de recherche et des rivalités. L'analyse pluraliste que nous avons esquissée (pluralité des approches et des objectifs de l'évaluation) soulève la question des procédures qui, impliquant tous ces acteurs, les aideraient à s'écouter les uns les autres, à délibérer sur les méthodes et critères d'évaluation, à atteindre des compromis et à prendre conscience du besoin de réviser périodiquement ces compromis. A quelles conclusions (ou plutôt interrogations) cette analyse conduit-elle ?

Les différenciations (qui restent élémentaires) que nous avons introduites pour rendre compte des complexités du processus de recherche conduisent à une première conclusion. L'emploi d'une batterie homogène d'indicateurs de performance pour évaluer la recherche, définir ses grandes orientations et la piloter dans tous les domaines serait au mieux une illusion, au pire un danger pour l'avenir d'un potentiel national de recherche. Car la compétition sur les moyens qui se déclenche porte sur un benchmarking trop abstrait et trop loin des contenus et dynamiques de recherches concrètes pour être d'un quelconque secours pour l'objectif principal, améliorer la qualité et la pertinence des travaux. On risque de n'avoir que des « tableaux qui tuent ». Il faut évaluer, quantifier, sélectionner le meilleur, refuser le mauvais, quelque soient les méthodes ou les critères, c'est évident. Mais les approches sont plurielles,

« Peer review and bibliometrics », Conference "Peer Review: Its Present and Future State" Prague, 12 – 13

de même que les acteurs pertinents de l'évaluation. Les communautés de recherche – qui, rappelons-le, sont des communautés bien souvent internationales - sont des acteurs nécessaires à tous les niveaux. Ces communautés de recherche existent, non seulement au niveau des grandes disciplines scientifiques, mais aussi au niveau des approches au sein d'une discipline ou à la frontière de plusieurs d'entre elles. Elles pratiquent, selon des pondérations variées, aussi bien le jugement par les pairs que l'emploi de batteries d'indicateurs, indicateurs qui jusqu'à un certain point diffèrent d'une communauté à l'autre.

La seconde conclusion est que la gouvernance de l'évaluation doit être posée comme problème. Celui-ci pourrait être résumé de la façon suivante.

a) La coexistence d'approches multiples doit être reconnue et aménagée. Le savoir que chacune apporte est partiel, mais doit être pris en compte dans le processus d'évaluation. Aucune ne peut prétendre détenir la vérité à elle seule. Il faut donc créer (ou conserver celles qui existent, si besoin en les ajustant) des procédures de délibération entre acteurs[17]. Elles sont d'une nature particulière, une de leurs fonctions principaales devant être de devenir des « sense-making fora »[18] (des forums créateurs de sens). Il s'agit en effet pour les participants d'arriver à une interprétation et une validation communes des données, de format et de nature diverses, produites par les différentes méthodes d'évaluation. C'est dans cet effort de créer un jugement commun que va se trouver mise en œuvre la satisfaction conjointe des quatre objectifs (contrôle, monitoring, apprentissage et développement). La délibération au sein de ce type d'instance est doublement importante. D'une part, chaque participant doit exprimer son point de vue aux autres. Le faisant devant ce qui constitue un public, il doit formuler ses arguments sous une forme recevable, c'est-à-dire s'appuyer sur sa conception d'une bonne recherche, sur des principes d'évaluation. D'autre part, arriver à un jugement commun n'interdit pas la permanence des désaccords, mais la sélection faite ensuite (par un vote ou toute autre procédure) l'est sur des bases correctes. Ces deux caractéristiques favorisent l'apprentissage de ce qu'est, dans sa diversité, une recherche de qualité.

October 2006, **(**http://www.pragueforscience.cz/Presentations.php).

17 Si on peut espérer une convergence des evaluations (par exemple, entre jugement par les pairs et mesure par indicateurs de performances) sur les dossiers très excellents ou, au contraire, très mauvais, le gros du travail et sa difficulté résident dans le choix parmi les bons dossiers, là où les divergences sont les plus fréquentes entre les méthodes et entre les experts.

b) De telles procédures d'évaluation sont situées, et même doublement situées : au niveau de chaque communauté de recherche et au niveau de chaque projet (projet de recherche individuel ou collectif, laboratoire). Respecter la diversité des critères de qualité suppose de placer les procédures au niveau des communautés pertinentes et d'y recruter des évaluateurs, nationaux et internationaux. L'évaluation interdisciplinaire devrait se faire en anticipant sur des communautés en devenir, par exemple en cours de structuration par des programmes de recherche. Comme c'est au travers de la rédaction du projet que peuvent être évaluées la qualité, la cohérence, la pertinence de la démarche proposée, la lecture des projets par tous les participants (ou par un panel) devrait être la règle idéale. Plutôt que d'évaluer des évaluations, il est bon parfois de s'intéresser directement au contenu de ce qui est évalué. Ceux qui sont évaluateurs dans les procédures européennes où chaque projet est évalué de manière indépendante par plusieurs experts qui se réunissent ensuite pour confronter leurs appréciations quantitatives et qualitatives, découvrent souvent qu'un jugement interactif est plus solidement fondé, scientifiquement parlant, qu'une introspection solitaire.

c) Pour que la délibération soit efficiente (sachant que toute façon il faut arriver à un résultat final) l'équilibre des voix (voices) doit être assuré entre les participants de manière à ce qu'ils soient entendus, alors qu'ils sont au départ inégaux en termes de pouvoir, de ressources, de compétences argumentatives et autres facteurs. C'est un problème classique en démocratie délibérative[19], redoublé dans le cas présent par la nécessité de réfléchir à l'organisation probable de plusieurs niveaux et lieux d'évaluation. Les objectifs, pluriels, de l'évaluation ne sont pas aisément compatibles entre eux. Une chose est de disposer d'arguments quantitatifs (relatifs à la performance) dans une négociation budgétaire où ces arguments sont dominants ; une autre chose est d'avoir un processus d'évaluation apte à améliorer sur le fond la qualité des recherches entreprises. Il faut donc imaginer des dispositifs à plusieurs niveaux avec des passerelles, mais aussi parfois des cloisons étanches.

d) Reste la question de l'innovation conceptuelle, méthodologique, empirique. Comment éviter la routinisation des procédures et de la sélection, un danger d'autant plus menaçant

[18] Selon l'expression de Uphoff, N. and Combs, J., 2001, *Some Things Can't be True But Are : Rice, Rickets and What Else : Conventional Wisdoms to Remove Paradigm Blockages*, Cornell International Institute for Food, Agriculture and Development, New York, Cornell University (cite par Engel et al., 2004) .

qu'on s'en tient à des méthodes standardisées, et ouvrir sur la nouveauté ? Toute innovation est dominée à ses débuts, car elle n'entre pas dans les références établies. Sa chance d'être bien évaluée dans un système d'indicateurs est dans l'ensemble plus faible, par exemple si elle se situe à la frontière de plusieurs disciplines. Elle demande une vision imaginatrice et ouverte des objectifs et des procédures.

[19] Voir, par exemple, Bohman, J., 1996, *Public Deliberation: Pluralism, Complexity and Democracy*, Cambridge MA, The MIT Press.