



**Conseil Economique
et Social**

Distr.
GENERALE

CES/AC.71/2001/4
30 octobre 2000

FRANÇAIS
Original: FRANÇAIS et
ANGLAIS seulement

**COMMISSION DE STATISTIQUE et
COMMISSION ÉCONOMIQUE POUR L'EUROPE**

**L'OFFICE STATISTIQUE DES
COMMUNAUTÉS EUROPÉENNES
(EUROSTAT)**

CONFÉRENCE DES STATISTICIENS EUROPÉENS

Réunion commune CEE/EUROSTAT sur la gestion de la technologie de l'information statistique
(Genève, Suisse, 14-16 février 2001)

Point (i): impact de l'entreposage de données sur la gestion des services de statistique

**UNE NOUVELLE ARCHITECTURE POUR LES SYSTÈMES D'INFORMATION
STATISTIQUE À EUROSTAT**

Rapport envoyé par Eurostat ¹

DOCUMENT SOLLICITÉ

I. INTRODUCTION

1. Eurostat, l'office statistique des Communautés européennes, gère actuellement plus d'une centaine de systèmes informatiques. Ces systèmes contiennent des données et des métadonnées statistiques qui viennent essentiellement des Etats membres et qui sont transformées pour être mises à disposition, sous forme harmonisée et souvent agrégée, d'utilisateurs internes, des institutions européennes, des Etats membres et des citoyens en général. Les opérations couvertes par ces systèmes sont diverses et nombreuses : réception des données, vérifications, estimations, harmonisation, transformations diverses, analyse, dissémination, pour mentionner les principales.

2. Pour des raisons d'économie et d'interopérabilité, il est impératif de coordonner étroitement le

¹ Préparé par Daniel Defays.

développement des systèmes et leur utilisation. Cette coordination s'exerce, entre autres, à travers la définition d'une architecture générale des S.I. Eurostat a ressenti le besoin, récemment, de faire évoluer cette architecture. Ce document présente ce qui motive ces nouveaux développements, les principes généraux qu'ils devront respecter, les principaux éléments de l'architecture et la stratégie de migration envisagée.

3. Ce projet est en phase de lancement. Ce qui est présenté n'est donc pas le compte rendu d'une évolution réussie, mais un plan encore en discussion et non encore validé lorsque ce document est écrit. Ces propositions n'engagent donc que l'auteur.

4. Il est certain que les instituts de statistique sont confrontés pour la plupart à des problèmes similaires et que des échanges sur des solutions envisagées ou apportées sont souhaitables. Les participants à la réunion du MSIT sont donc invités à commenter ces propositions à la lumière de leurs expériences personnelles.

II. CE QUI MOTIVE L'EVOLUTION DE L'ARCHITECTURE

5. L'architecture actuelle est structurée en environnements. Un environnement est, dans le jargon d'Eurostat, défini à partir d'un ensemble de données, d'activités/fonctions qui lui sont liées, et d'acteurs correspondants. Trois niveaux sont distingués : la production, la référence et la diffusion. En amont, des outils de collecte standards sont mis à disposition des utilisateurs. L'organisation de la production est décentralisée; alors que pour la référence et la dissémination des unités spécifiques exerçant une coordination forte existent.

6. Ce modèle doit évoluer pour différentes raisons :

- ◆ Le mode de gestion décentralisé de la production atteint ses limites; les coûts associés doivent mieux être maîtrisés grâce, entre autres, à une meilleure coordination qui doit être garantie par l'architecture.
- ◆ Le volume de plus en plus important de données confidentielles à traiter et le souci croissant de protection de leur confidentialité nécessitent de repenser le traitement de ces données et de revoir la notion d'environnement de production en conséquence.
- ◆ Les données reçues par Eurostat ne font pas actuellement l'objet d'un archivage systématique coordonné; il est pénible de refaire tourner une chaîne de traitement sur des données originelles en cas de problème.
- ◆ Les nécessités croissantes de documenter les données et de leur associer des méta informations pour faciliter leur traitement et leur accès amènent les statisticiens à traiter de plus en plus de données non numériques; le rôle et la place de celles-ci dans l'architecture actuelle ne sont pas suffisamment précisés.
- ◆ Le mode de conception linéaire de la chaîne statistique (collecte, production, diffusion) sur lequel est basé l'architecture n'est pas toujours adéquat : production partagée, mise à disposition par Eurostat de données venant d'autres sources sans traitement interne, diffusion de métadonnées. Internet contribue aussi à faire éclater cette vision linéaire et quelquefois fermée du traitement statistique. De plus, les

applications administratives, de plus en plus nombreuses, n'épousent pas non plus le découpage.

- ◆ La version actuelle de l'environnement de référence est plus proche des besoins de la production que ceux de la diffusion : les producteurs y retrouvent facilement leurs données tandis que les clients ont souvent beaucoup de difficultés à y retrouver les données qu'ils cherchent. Cela est dû à l'insuffisance de métadonnées d'aide à la recherche et à l'absence d'un modèle de présentation des données unique et orienté clients.

III. QUELQUES CARACTERISTIQUES SOUHAITABLES DE LA NOUVELLE ARCHITECTURE CIBLE

7. Une analyse du cycle de vie des données à Eurostat et des problèmes rencontrés avec l'organisation actuelle a débouché sur les constats suivants.

8. Il existe essentiellement quatre ensembles de données stratégiques. Ces ensembles ont un rôle structurant sur l'architecture des S.I. et l'organisation d'Eurostat:

- ◆ données et métadonnées brutes venant des fournisseurs d'informations;
- ◆ données nettoyées et harmonisées qui constituent les éléments de base à partir desquels sont élaborés tous les produits statistiques; ces données, dites de référence interne, sont éventuellement confidentielles;
- ◆ données de référence externe qui peuvent être consultées par les utilisateurs;
- ◆ données diffusées.

9. La production des données est fortement décentralisée. L'architecture doit épouser ce mode d'organisation et laisser un maximum de souplesse aux utilisateurs. Par contre, il existe des phases où le processus de production est plus fortement coordonné (réception des données à travers Stadium/Statel, mise à disposition des données dans la base Comext et NewCronos). Ces phases correspondent aux interfaces d'Eurostat avec l'extérieur. Pour des raisons de cohérence et d'efficacité, cette centralisation actuelle doit subsister.

10. Le concept d'environnement sécurisé doit être adapté de manière à permettre de travailler de manière plus confortable sur des collections de données qui contiennent des informations partiellement confidentielles. Idéalement, une bonne partie de l'environnement de production devrait se situer en zone sécurisée.

11. L'architecture doit être conçue de manière à favoriser l'interopérabilité et la réutilisation des différents systèmes ou applications. Ceci permettra de travailler économiquement de manière décentralisée. Cette interopérabilité sera garantie en définissant des interfaces standards et des mécanismes d'échanges appropriés entre systèmes d'information.

12. Des métadonnées standardisées doivent permettre de garantir la cohérence des données et de les documenter. Elles seront définies à partir des besoins exprimés par les clients, les producteurs, en remontant ensuite la chaîne jusqu'aux fournisseurs d'informations.

13. De plus, il apparaît important, pour des raisons d'efficacité, de fiabilité et d'économie des ressources:

- ◆ de n'adopter que des solutions testées, qui ont fait leur preuve dans d'autres environnements;
- ◆ de prévoir une évolution graduelle qui s'inspirera étroitement de la situation actuelle.

Cette dernière exigence s'avérera avoir un rôle fort contraignant dans la suite des investigations.

IV. BENEFICES POUR LES PRODUCTEURS DE DONNEES ET LES UTILISATEURS

14. Un chantier aussi important que la révision de l'architecture générale des S.I. ne peut être couvert que s'il est raisonnable d'en attendre des bénéfices substantiels pour l'institution elle-même, ses producteurs de données et ses clients.

15. Le nouveau modèle doit offrir:

- ◆ une rationalisation importante des outils informatiques;
- ◆ des outils mieux ciblés aux besoins des différents groupes de clients : les utilisateurs internes à Eurostat, les directions générales de la Commission, les fournisseurs externes (INS ou autres), les autres clients (spécialistes ou généralistes);
- ◆ un meilleur accès aux métadonnées et une meilleure articulation des bases de données;
- ◆ une plus grande traçabilité des traitements effectués;
- ◆ un plus grand confort dans le traitement des données et particulièrement des données confidentielles.

V. LES QUATRE ENVIRONNEMENTS CONSTITUTIFS DE LA NOUVELLE ARCHITECTURE

16. L'analyse du cycle de vie des données, le recensement des opérations effectuées sur les données et la définition du concept d'environnement induisent une structuration en 4 niveaux légèrement différente de celle existante. Il importe de remarquer que les environnements ne sont pas nécessairement des constructions physiques différentes.

V.1 Environnement de réception

17. Il s'organise autour du répertoire des données et métadonnées fournis par nos correspondants nationaux. Ce "production data repository" constitue en quelque sorte le patrimoine d'Eurostat et le point de départ de tous les traitements qui seront effectués. A cet environnement, sont rattachées les activités de "data capture", "validation", "error correction".

V.2 Environnement de production

18. Il s'organise autour de la référence interne et constitue la source unique d'alimentation des environnements de référence et de dissémination. Il est le lieu de dépôt des données primaires, validées et corrigées, ainsi que celui de la construction d'estimations et de données dérivées par combinaison de données de différents domaines, le cas échéant. Il n'est accessible qu'au personnel d'Eurostat.

19. Cet environnement contient l'ensemble de microdonnées, macrodonnées d'Eurostat, portions confidentielles incluses. Il contient également l'ensemble des métadonnées nécessaires à la production ainsi que celles nécessaires aux environnements de référence externe et dissémination. (Dictionnaires, nomenclatures, mots clés, libellés multilingues, footnotes, notes méthodologiques, relations diverses en objets statistiques, formules, ...)

20. A cet environnement, sont rattachées les activités de "transformations and derivations", "estimation", "data inspection and editing", "statistical analysis", "nomenclature preparation and house keeping".

V.3 Environnement de référence

21. Il s'organise autour de la référence externe et contient des données et métadonnées de qualité et non confidentielles, qui peuvent être diffusées à l'extérieur d'Eurostat. Il est accessible à travers une interface unique.

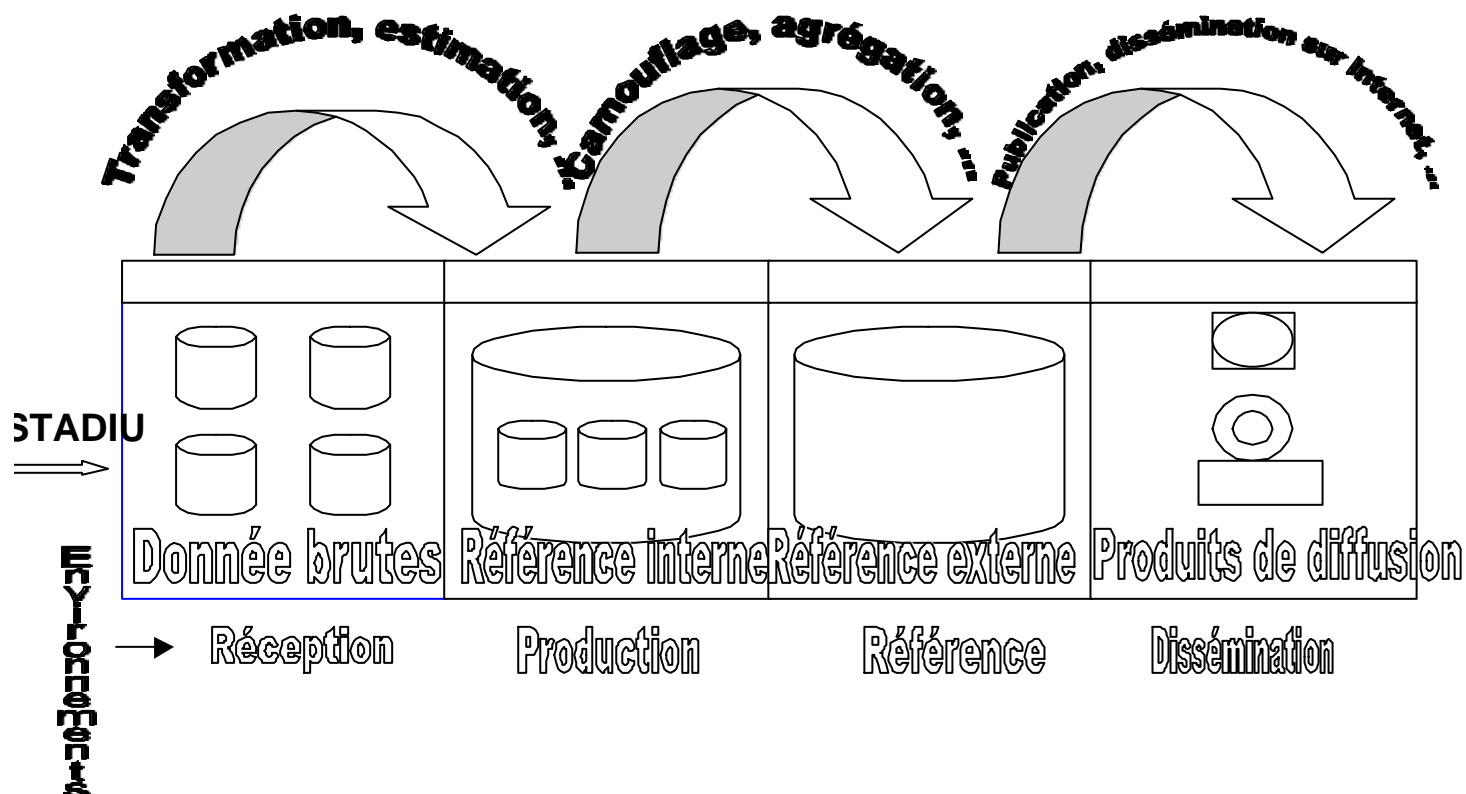
22. Le public concerné est l'ensemble des personnes, organismes, intéressés de manière systématique par le détail des données statistiques; institutions européennes et internationales, ministères, data shops, organismes fournisseurs de données.

23. A cet environnement, sont attachées les activités de "camouflage", "aggregation", "application of flags and footnotes", "supply of data to reference DB's", "nomenclature preparation and house keeping", "statistical analysis".

V.4 Environnement de dissémination

24. Il s'organise autour d'une version électronique de toutes les données diffusées par Eurostat. Ces données constituent la vitrine d'Eurostat. A cet environnement sont rattachées les activités de "preparation for publication" et "dissemination".

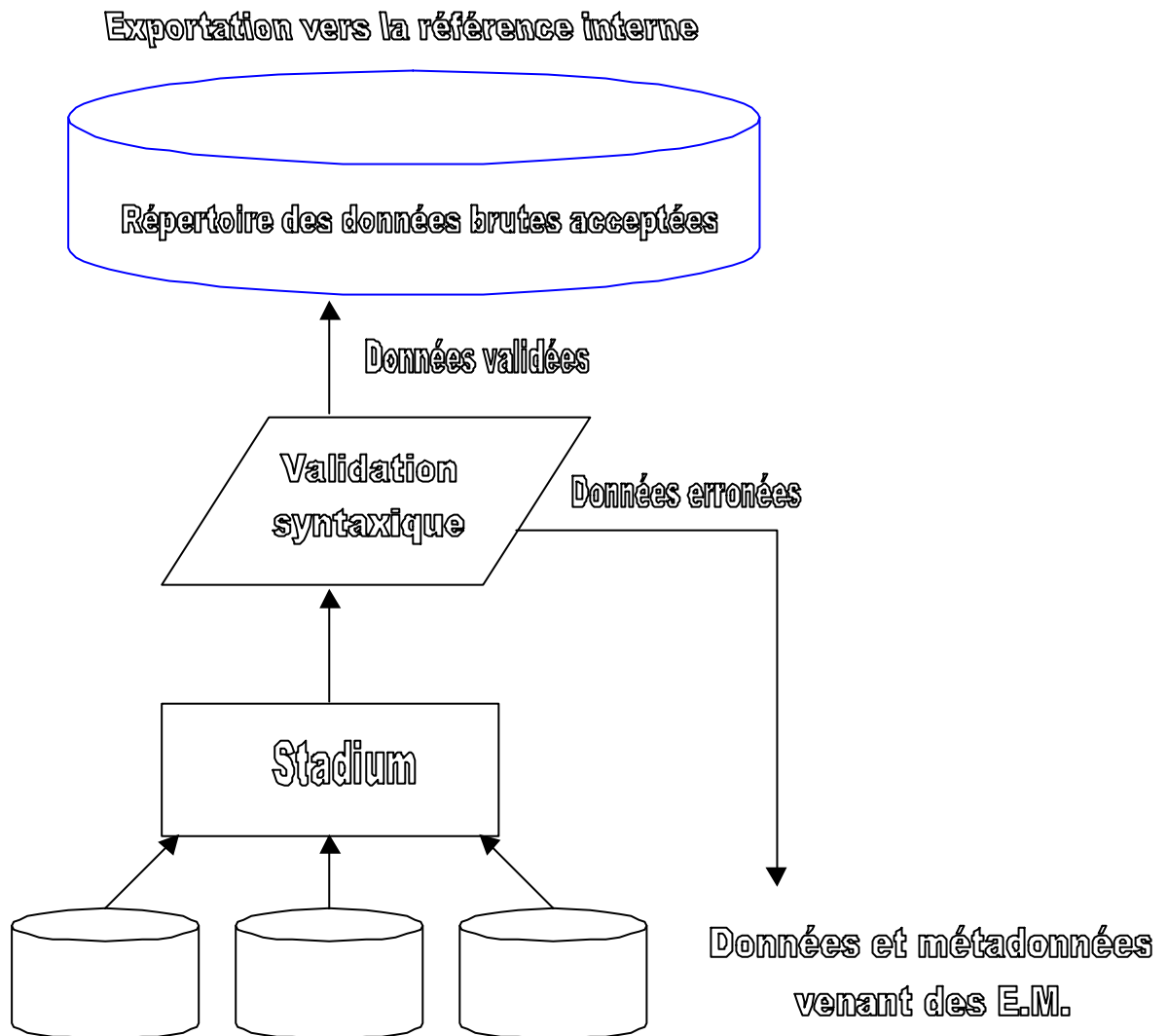
Graphique 1. Les environnements constitutifs de l'architecture



VI. ELEMENTS DE L'INFRASTRUCTURE LOGICIELLE

25. Chacun des environnements décrits ci-dessus se matérialise dans des bases de données (réelles ou virtuelles), des boîtes à outils, des composantes logicielles, etc. Ce paragraphe décrit la situation cible et l'état actuel des systèmes.

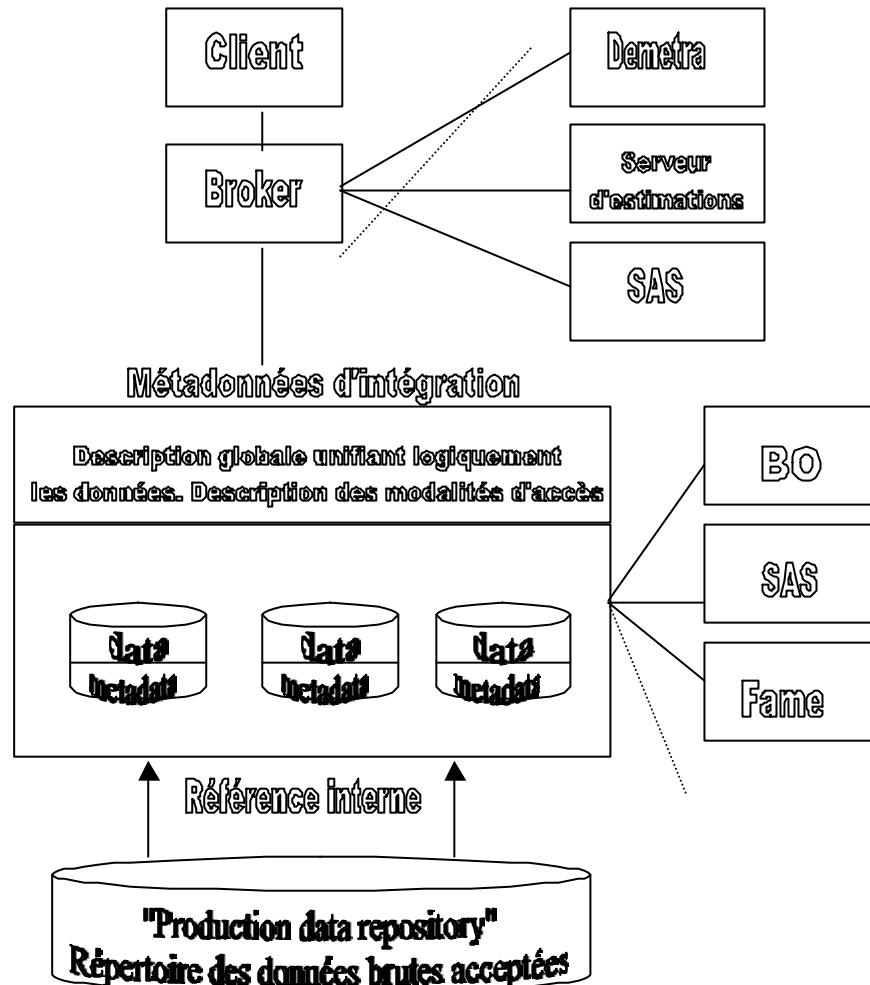
VI.1 Environnement de réception



26. Cet environnement s'organise autour d'outils de réception/enregistrement des données existants (Statel, Ediflow et Stadium). Les données sont envoyées en utilisant GESMES. Le producteur réceptionne les informations transmises via l'outil Stadium et s'assure de leur conformité syntaxique. Le cas échéant, il renvoie le fichier à l'expéditeur. Une fois validées, les données sont reconnues par Eurostat et par

l'expéditeur comme étant le point d'entrée de la chaîne de traitements. Elles sont archivées. Actuellement, cet archivage a la forme d'un pointeur vers un fichier directement sous le contrôle du producteur. A terme, un stockage sous un seul DBMS serait souhaitable pour des raisons de sauvegarde et d'interopérabilité.

VI.2 Environnement de production



27. L'environnement de production, et plus particulièrement la référence interne, doit offrir des possibilités de communication interdomaine, d'accès aisé et sécurisé par les logiciels de traitement externes.

Il sera conçu en tirant parti de la situation existante, c'est-à-dire en ajoutant une couche descriptive qui permette une intégration virtuelle des données. Cette couche, appelée "métadonnées d'intégration" contiendra, entre autres:

- ◆ des descriptions de formes pivots et des correspondances de noms, pour chaque type de données accessibles;
- ◆ des pointeurs d'objets physiques vers des objets composites;
- ◆ des descriptions des types et formats physiques des données accessibles, y inclus le type de logiciel de stockage;
- ◆ des informations sur la localisation des données, des éléments pour permettre des interrogations

harmonisées (dimensions, valeurs standards, ...);

- ◆ la description du comportement de base du logiciel interrogé (erreurs de retour, etc.);
- ◆ le nom des fonctions statistiques, leurs paramètres d'entrée, les types de résultats ainsi que la localisation du programme applicatif.

Il s'agit très clairement de la partie névralgique de cet environnement.

28. De plus, les fonctions seront, dans la mesure du possible, encapsulées dans des outils logiciels réutilisables. Les opérations devraient se dérouler comme suit: l'utilisateur accède à l'information contenue dans la référence interne (c'est-à-dire stockée physiquement dans les bases SAS, Oracle, Fame, ...) via une composante client, qui aurait la capacité de construire des requêtes, lancer des analyses qui impliqueraient différents serveurs applicatifs, et parcourir des résultats. Le client communique les clauses d'interrogation au "broker", pièce logicielle frontale dont le travail essentiel est de recevoir des commandes, de les acheminer vers les serveurs applicatifs ou les serveurs de données ainsi que de recevoir des messages sur l'état des travaux, pour prendre ensuite les décisions adéquates.

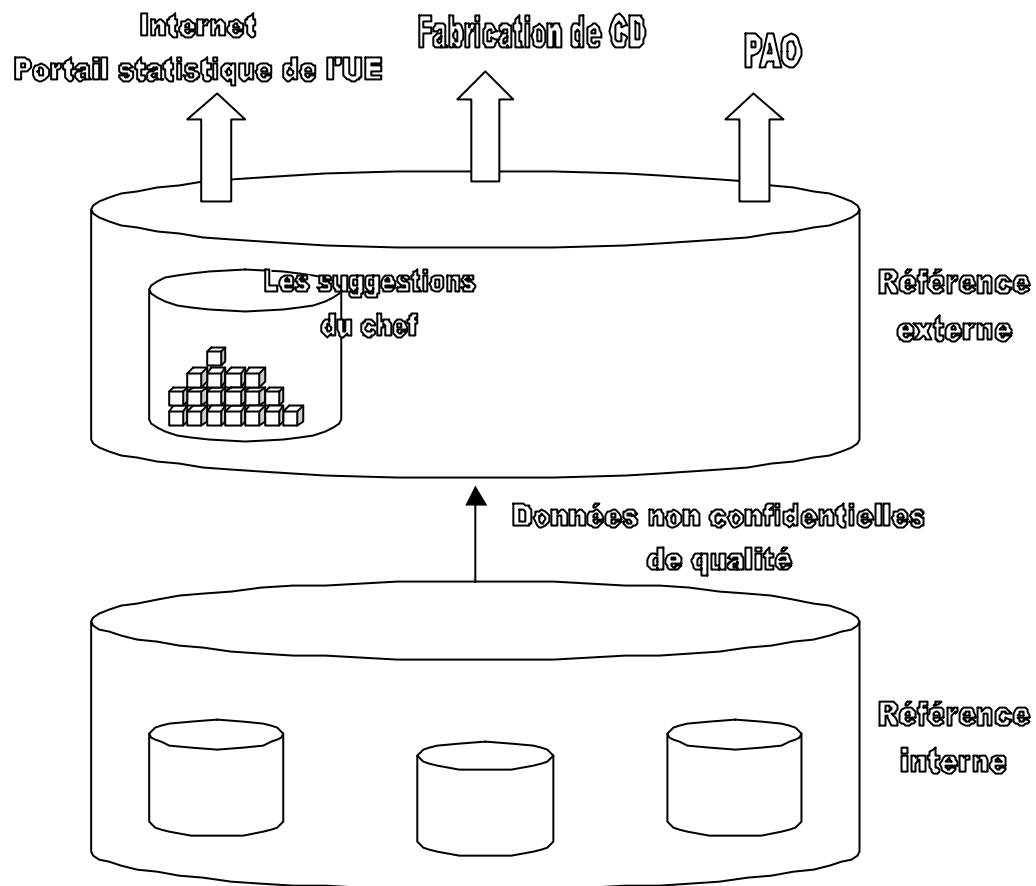
29. Les traitements génériques se font à travers des composants réutilisables qui prennent la forme de ce qui a été appelé des serveurs applicatifs (Demetra, serveurs d'estimation, applications SAS dans le graphique). Un serveur applicatif est un logiciel, fréquemment sans mémoire, utilisable séparément. De manière à permettre d'enchaîner les traitements, les entrées et sorties sont standardisées.

30. Il est également envisagé de garder des logiciels de type alternatif (à transactions ouvertes - sans retour à la source) qui n'utilisent pas le broker. Il s'agit de produits de type SAS, Oracle Express, Business Objects, Fame, TPL ou ACCESS/EXCEL. Une caractéristique commune de ces logiciels est l'usage de SQL et la création de leur propre repository. Eux aussi constituent une pièce fondamentale de l'environnement de production. Ils sont actuellement largement utilisés à des buts d'analyse mais aussi de réalisation de fonctions complexes relevant de l'environnement de production (indices, échantillonnage, ...), et paraissent irremplaçables.

31. Le concept "métadonnées d'intégration" a été testé dans l'application Comext. Il existe actuellement une interface qui permet d'unifier les données de la principale base de référence d'Eurostat, NewCronos et du commerce extérieur en utilisant le principe proposé dans ce document.

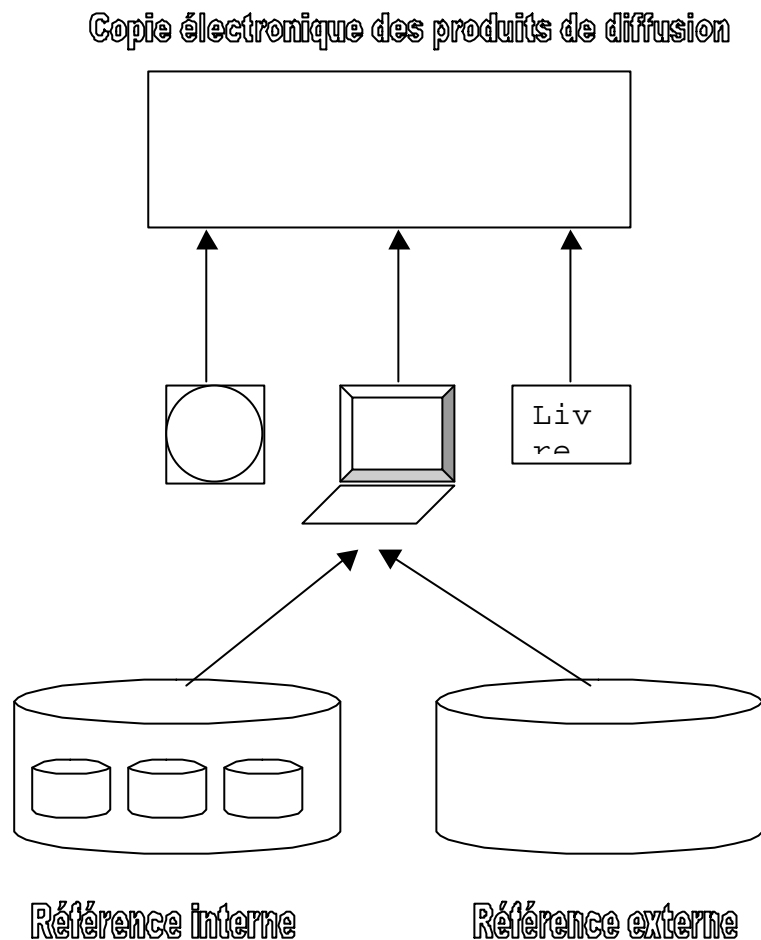
32. Des embryons de serveurs applicatifs existent : Demetra pour la désaisonnalisation, un serveur de camouflage CIF, un logiciel de préparation des données (DPS) écrit en SAS. De plus, l'architecture basée sur un broker va être testée dans le cadre du développement d'un serveur d'estimation.

VI.3 Environnement de référence



33. Cet environnement doit permettre une présentation standardisée des données, un contrôle stricte des accès, et fournir une interface unique et un seul langage pour réaliser des extractions et stocker des données. Pour des raisons d'accès et de robustesse de l'ensemble de l'architecture, il est proposé d'en faire un environnement physiquement séparé (au prix d'une redondance des données ...). Il est construit autour de la base de référence existante NewCronos ou de son successeur. Aucun changement majeur ne paraît nécessaire à court terme si ce n'est l'uniformisation des interfaces (Comext/NewCronos). A terme, il importe de se libérer du format actuel de stockage qui est propriétaire et rend les références croisées assez lourdes.

VI.4 Environnement de dissémination



34. Les données diffusées devront satisfaire à des standards de qualité. La gestion des accès sera élaborée, avec des portails correspondants à différents types de clients. Des facilités interactives d'abonnement, de notification automatique, de personnalisation des accès seront développées. "Eurostat site 2" constitue actuellement la partie Internet de cet environnement et des évolutions soutenues par les avancées de l'e-commerce sont à attendre dans les années à venir.

35. Le changement majeur par rapport à la situation actuelle viendra de la nécessité de garder une copie électronique, archivée de manière centralisée, de l'ensemble des produits diffusés.

VII. STRATEGIE DE MIGRATION

36. Faire évoluer une architecture, en respectant les exigences de la production, à l'intérieur d'un budget constant est un exercice périlleux. Ces deux contraintes ont, du reste, contribué à limiter les ambitions. Le nouveau modèle sera implémenté de manière incrémentale, à partir de modules testés et

robustes, comme cela a déjà été mentionné.

37. La stratégie adoptée repose sur la démarche suivante :

- ◆ Association large des utilisateurs avertis à la définition du projet. Un document d'orientation générale décrivant ce qui motive le projet et ses caractéristiques principales a été débattu au sein d'Eurostat.
- ◆ Implication de la hiérarchie. Le comité de direction a donné son accord sur l'esprit du projet et sur une première analyse du cycle de données à Eurostat.
- ◆ Etude détaillée des objets statistiques et des fonctionnalités requises. Une analyse fouillée des différents types de données et métadonnées manipulées à Eurostat, s'inspirant entre autres des travaux de B. Sundgren et de METIS, a été réalisée. En parallèle, les opérations subies par ces données au cours de leur cycle de vie ont été répertoriées.
- ◆ Première rationalisation des outils utilisés pour la production. La diversité des logiciels utilisés à Eurostat et des applications spécifiques a permis aux statisticiens de rencontrer de manière souple et rapide la majorité de leurs besoins. Le coût pour l'institution est cependant fort élevé : licences, formation, maintenance. Parallèlement à la mise en place de la nouvelle architecture, il a donc été décidé de limiter le nombre d'outils de développement et de réaliser au maximum des modules existants. Ceci devrait simplifier l'environnement de production existant et faciliter la migration.
- ◆ Construction de prototypes. Certains éléments clés de l'architecture, comme la couche "métadonnées d'intégration", le "broker", les serveurs applicatifs font actuellement l'objet de prototypes.
- ◆ Parangonage. Des systèmes d'information de certains Etats membres et des recommandations internationales ont été étudiés de manière à tirer parti des expériences acquises par ailleurs.
- ◆ Approche incrémentale. L'architecture proposée permet une approche incrémentale. Dans un premier temps on se concentrera sur les systèmes d'information principaux, tournant actuellement sous FAME, ACUMEN, Oracle/Express et SAS.

Les différentes applications seront intégrées progressivement.

38. Un des principaux intérêts de la solution envisagée est de se superposer au dispositif existant, essentiellement. Ceci donne beaucoup de marge de manœuvre dans le déploiement du nouveau modèle.

VIII. CONCLUSIONS

39. Cet article a fait l'état de réflexions au sein d'Eurostat sur le développement d'une nouvelle architecture des systèmes d'information statistiques. Le succès d'une opération de cette envergure dépendra non seulement de la pertinence du modèle proposé mais également de la capacité de l'institution à changer. Les comportements devront s'adapter, et ceci n'est pas le moindre défi.

40. Dans la solution envisagée, le souci majeur a été de conserver un équilibre entre une informatique au service des utilisateurs, capable de réagir rapidement à de nouveaux besoins, à des nouvelles contraintes, grâce à une grande autonomie des utilisateurs dans le développement et l'utilisation des outils nécessaires, et une informatique plus corporate, qui utilise autant que faire se peut des solutions génériques, et qui garantit l'interopérabilité des systèmes locaux.