Une histoire de S à l'I.N.R.A.

Guy FAYET

Laboratoire d'Informatique et Génie Logiciel
Champenoux - 54280 - SEICHAMPS
fayet@nancy.inra.fr

Résumé : en complément de l'article de Carlier présentant le logiciel S, on raconte comment depuis quelques années ce logiciel a été pris en charge à l'I.N.R.A., les espoirs qu'on y porte et les qualités qu'on lui trouve.

Mots-Clés : environnement statistique, langage statistique, objets statistiques, extensibilité, qualité.

Plutôt que d'ajouter des détails techniques à l'excellente présentation de S disponible dans ce numéro de la Revue de Modulad, ce papier cherche à raconter - en en tirant les principaux enseignements - l'histoire du logiciel S dans l'I.N.R.A., à travers trois étapes principales :

- pourquoi et comment
- les travaux de première période
- la généralisation dans l'Institut

Au travers de ce récit, on espère apporter un témoignage utile sur les enjeux, les difficultés, mais aussi les enthousiasmes et certaines réussites, liés à une rationalisation et une modernisation de l'atelier logiciel dont dispose le chercheur biologiste

On concluera par des considérations qui restent, bien entendu, l'expression de la seule opinion de l'auteur

1. Pourquoi et comment : le choix de S

1.1. L'I.N.R.A. en bref

Constitué de 22 Centres de recherche répartis sur le territoire métropolitain, l'I.N.R.A. (Institut National de la Recherche Agronomique) regroupe environ 8300 personnes, dont 2500 chercheurs et ingénieurs. Une cinquantaine de Domaines Expérimentaux et Laboratoires isolés complètent le dispositif

Au plan informatique, trois niveaux sont utilisés conjointement jusqu'en 1988 : un serveur central (DPS 8 sous Multics), une vingtaine de Mini 6 et DPS 6 de Bull répartis dans les Centres, et plus de 2000 micro-ordinateurs de toutes variétés. Les logiciels employés dans le domaine du traitement et de la représentation des données sont très variés, peu de passerelles existent entre eux, la plupart appartiennent à la génération "d'avant l'interactivité".

1.2. Un schéma directeur

La disparition annoncée du système Multics, la faible puissance des serveurs locaux, l'inadéquation de la panoplie des logiciels aux besoins des scientifiques, et l'écart entre la technologie employée et celle disponible sur le marché conduisent en 1987 à la décision d'élaborer un nouveau Schéma Directeur de l'Informatique Scientique (S.D.I.S.)

Plusieurs années seront nécessaires à son élaboration et sa mise en place (encore en cours aujourd'hui). Les principales options sont cependant prises assez tôt.

1.2.1. Matériels et réseaux

Assez classiquement, choix est fait d'augmenter considérablement la puissance disponible en la répartissant - selon les besoins locaux - en serveurs de Centres, serveurs de Laboratoires, postes de travail. Un poste de travail est un terminal, un

micro-ordinateur sous MS-DOS[1], un terminal X-Window, ou une station de travail sous Unix[2] Tous les serveurs exploitent le système Unix

Tous les postes de travail fonctionnant sous Unix et tous les serveurs sont reliés en réseau IP[3], utilisant selon les cas les supports Ethernet, fibre optique, ou Transpac. L'ensemble de l'I N R A est ainsi interconnecté. De plus, une proportion croissante de micro-ordinateurs sont reliés à ces réseaux, soit par voie asynchrone, soit à l'aide d'une carte Ethernet additionnelle.

1.2.2. Un fil directeur : la boîte à outils

Les études du S.D.I.S montrent que, en dehors des domaines spécifiques où l'informatique est un matériel de laboratoire spécialisé de la thèmatique de recherche (par exemple les matériels et logiciels spécialisés dans l'analyse et la représentation des molécules), une très grande partie des besoins fonctionnels des chercheurs est commune à tous les domaines d'étude On peut citer notamment :

- la communication
- la bureautique
- les bases de données
- les systèmes d'information documentaire
- la manipulation et le traitement de données brutes ou élaborées
- les représentations graphiques
- les processus mathématiques (numériques) de simulation, d'optimisation, d'estimation, de modélisation
- la flexibilité : programmation, utilisation de bibliothèques, ...
- la personnalisation : gestion de l'écart entre le respect de normes[4] et la culture personnelle ou celle de l'équipe

Ainsi est né le concept de boîte à outils du scientifique, ensemble de fonctionnalités réalisées par une panoplie de logiciels communs à tout le monde Des économies d'échelle sont ainsi permises par des marchés portant sur de "grosses quantités".

1.2.3. Evaluation de S

Dans le domaine du traitement et de la représentation graphique de données numériques, fort nombreuses dans l'expérimentation agronomique, l'I N.R.A utilise traditionnellement des produits comme Amance, Modli, Modulad, Spad, Addad, SAS, StatITCF, StatGraphics, etc.

Une mention spéciale doit être faite du progiciel CS (Consistent System) qui était spécifiquement attaché au système Multics Il fallait trouver sous Unix un logiciel à

1 - MS-DOS est une marque déposée de Microsoft

2 - Unix est (le croiriez-vous ?) une marque déposée des Bell-Laboratories

3 - Internet Protocol : quelle que soit la couche physique employée, l'homogénéité d'accès et d'action est ainsi assurée aux utilisateurs

4 - de programmation, cela vient en premier à l'esprit, mais également - surtout ? - d'habitudes personnelles, ou d'équipe, ou de domaine d'action

peu près équivalent dans ses fonctionnalités. C'est ainsi qu'une évaluation de S a été conduite :

- sur le plan fonctionnel : interactivité, méthodes statistiques et structures de données disponibles, capacité à automatiser ou personnaliser des sessions (macros ou fonctions), possibilités graphiques.

- sur le plan informatique : fiabilité et précision des calculs, ressources utilisées, qualité du code, efficacité, portabilité

- sur le plan de l'interface utilisateur : langage de commande, documentation imprimée et en ligne, périphériques graphiques supportés, etc.

Cette évaluation a été conduite essentiellement par le département de Biométrie et le département d'Informatique. On imagine après la lecture de l'article de Carlier quelle fut la décision.

1.3. Obtention du produit

L'I.N.R.A. s'est donc procuré auprès de AT & T deux licences du source de S. Bien qu'il existe des distributeurs de versions binaires, le source nous était indispensable pour des raisons de portage, notamment sur des machines Bull (SPS 7 puis DPX 2000 et DPX 5000), spécialité nationale. En sus de ces licences, un droit de copie binaire nous permettait de distribuer les exécutables dans l'Institut.

Le mode d'établissement des prix peut être intéressant à connaître : la première licence source coûte[5] 8000 US$, la seconde 3000 US$. Le droit de distribution est de 5000 US$, augmenté à chaque installation de royalties dont le montant varie avec la taille de la machine cible (de 100 US$ pour une station de travail à 1400 US$ pour un gros serveur).

2. Première période : appropriation du logiciel

2.1. La fourniture

Très classiquement, le logiciel arrive sous forme de bande magnétique accompagnée d'un (petit) manuel d'aide à la génération et de la documentation utilisateur. L'étude du source montre une très grande structuration du produit, l'emploi de deux langages principaux : C pour le noyau de S (gestion du langage de commande, des structures de données, des pilotes graphiques, de la communication entre co-processus) et Fortran pour tous les algorithmes numériques.

En fait, les préprocesseurs[6] classiquement disponibles sous Unix sont massivement utilisés. La portabilité est ainsi rendue maximale. Certains paramètrages sont à

5 - Ces prix correspondent au tarif d'AT & T de l'époque, ils ne préjugent en rien des prix pratiqués par les distributeurs.

6 - Il s'agit de *m4*, *cpp*, *ratfor*. L'analyseur du langage de commandes est généré avec les outils *lex* et *yacc*.

opérer, notamment sur la représentation des nombres, et les passages de structures de données entre C et Fortran.

2.2. Les portages

La génération de cette version[7] sur divers matériels a montré un très petit nombre de bogues. En fait, la presque totalité des problèmes rencontrés au cours de plus d'une cinquantaine de portages provient d'anomalies dans les outils fournis par les constructeurs : préprocesseurs, compilateurs, bibliothèques, gestion des signaux[8].

Une opération de portage sur une nouvelle version de système représente une charge moyenne de 3 jours x hommes. Un changement d'architecture de machine est souvent plus lourd : de 5 à 30 jours x hommes.

2.3. Les extensions

L'évaluation préalable avait montré deux catégories d'insuffisance du produit natif :

- au niveau de la liste des pilotes graphiques disponibles ;
- au niveau de méthodes statistiques largement employées dans l'I N R A.

Des développements spécifiques ont donc été conduits dans ces deux domaines

2.3.1. Pilotes graphiques

Deux pilotes ont été développés : l'un pour des petits traceurs à quatre couleurs (Benson 1002) qui interprètent le code *hpgl* de Hewlett-Packard, l'autre pour des imprimantes matricielles monochromes ou polychromes de Data Products[9].

Ces pilotes ont été facilement développés par imitation des pilotes existants.

2.3.2. Modèle linéaire

C'était la plus grave carence de S. Le programme Modli a été repris[10], en le décomposant en fonctions élémentaires. Pour chaque étape, une fonction d'appel a été écrite en langage S, les analyses étant ainsi décomposées et tous les résultats, finaux ou intermédiaires, sont disponibles à l'utilisateur sous forme de structures de données S, sur lesquelles il peut faire intervenir l'ensemble du langage.

---

7 - Version de 1984, dite "Old S" depuis la nouvelle version qui date de 1989.

8 - Nous utilisons maintenant cette opération de génération de S pour tester la conformité des outils fournis par les constructeurs. Il est très rare qu'une anomalie - ou une spécificité - existe qui n'ait pas été mise en évidence ainsi.

9 - Depuis ont été ajoutés un pilote LaTeX, de production interne, et un pilote X11 "récupéré" dans le chapitre S de la bibliothèque publique `statlib`

10 - Guy DECOUX - Laboratoire de Biométrie - I N R A. - Versailles

2.3.3. Base de données

La boîte à outils contenant à la fois un logiciel de structuration et d'archivage de données (en l'occurence le SGBDR Oracle[11]) et un logiciel de manipulation et de traitement de données, il paraissait logique d'offrir une passerelle aux utilisateurs, leur permettant d'importer dans S une relation, ou d'exporter vers leur base de données une structure de données de S. Un prototype fut réalisé en 1989.

La version définitive, disponible dans les prochaines semaines, permet toute la puissance du langage SQL et repose sur un modèle client / serveur pouvant fonctionner à travers le réseau entre machines différentes.

Il faut noter que grâce à la qualité du code[12] et des bibliothèques fournies, ces développements n'ont pas posé de problème majeur. Tous s'articulent sur des co-processus, activés par une fonction spécifique ajoutée à S, et qui communiquent par signaux, tubes, mémoire partagée, etc.

3. Généralisation : la version de 1989

3.1. Particularismes

En 1989 sortait une nouvelle version de S, essentiellement caractérisée par une bien plus grande facilité d'extension : il n'existe plus de macros avec un langage spécifique, mais tout utilisateur peut développer ses propres fonctions directement en langage S lui-même. La plupart des fonctions natives sont d'ailleurs ainsi écrites, le noyau de S opérant ainsi une sorte de "boot" pour s'enrichir de ses propres fonctionnalités.

Une autre amélioration porte sur le langage lui-même : il permet une description "presqu'objet" des structures manipulées, attachant aux données des attributs et des évaluateurs. Des expressions particulières portent sur le langage, une fonction étant ainsi considérée comme une structure de donnée - un "objet" - banalisée. La puissance d'expression, la rigueur, la fiabilité des fonctions développées "à la maison" en sont accrues.

La génération et l'installation sur les plateformes de l'I.N.R.A. de cette nouvelle version n'ont pas posé de problèmes graves. L'incorporation des extensions propres à l'Institut s'est opérée rapidement. Cette version est actuellement installée sur une centaine de machines dans l'I.N.R.A.

3.2. Les nouvelles extensions

En 1990 se tenait à Wellington le premier atelier international des utilisateurs de S. Ce fut l'occasion d'une riche expérience et la découverte d'une large communauté très ouverte.

---

11 - Oracle est une marque déposée de Oracle Corp.

12 - Pas de la documentation : si la documentation utilisateur est de très bon niveau, il n'existe pas de documentation de réalisation. Il faut parfois aller "à la pêche" dans le source.

La concrétisation en est la disponibilité de nombreuses extensions et améliorations, à travers une conférence spécialisée[13] et une banque de produits[14]. On y trouve de nouvelles fonctionnalités statistiques et des passerelles vers d'autres logiciels[15].

La mise en ligne d'une nouvelle extension est particulièrement simple : s'il ne s'agit que de fonctions écrites en langage S, cela est immédiat (une commande). Si l'apport contient des modules écrits en C ou Fortran, une fonction d'appel et de passage d'arguments est disponible pour les activer. Les bibliothèques d'objets étant fournies, il suffit de refaire une édition des liens. Pour les architectures qui le supportent, il existe même une solution de chargement dynamique, qui n'encombre ainsi la mémoire de la machine que lorsque les modules externes sont réellement appelés par l'utilisateur.

Lors de la prochaine diffusion interne, S sera enrichi (outre les extensions déjà citées) d'une documentation en ligne en français, d'une bibliothèque de fonctions traitant du modèle non-linéaire[16], d'une autre bibliothèque traitant de statistique non-paramètrique[17], et d'un maximum d'extensions issues de *statlib*

3.3. L'environnement mis en place

Introduire ainsi un progiciel dans une communauté d'utilisateurs doit s'accompagner d'un certain nombre de mesures d'aide et de conseil

Ainsi plusieurs cours sont dispensés, plus ou moins spécialisés en fonction du public, offrant aux utilisateurs une "mise à l'étrier" autour du langage S et des fonctionnalités disponibles. Le logiciel est également utilisé, parmi d'autres, comme support informatique dans un enseignement interne de statistique

Une organisation de conférence électronique (*news* sous Unix) est mise en place pour assurer l'aide à distance, chacun pouvant intervenir comme questionneur ou répondant. Notre laboratoire est responsable de cet aspect conseil, et de la maintenance en général. Le département de Biométrie renseigne sur les aspects de méthodologie statistique

3.4. Des projets pour le futur

Nous envisageons le développement d'un certain nombre d'interfaces entre S et d'autres progiciels de la boîte à outils :

13 - `S-news@stat.wisc.edu`. On peut se faire inscrire dans la liste d'envoi en écrivant à `S-news-request@stat.wisc.edu`.

14 - `statlib@stat.wisc.edu`

15 - Par exemple vers *SAS* de SAS Institute, *emacs*, *Frame Maker* de Frame Technology Corp., etc.

16 - *NL*, produit issu des travaux du département de Biométrie

17 - Travaux de M. TERQUI - Laboratoire de Physiologie de la Reproduction - I.N.R.A. - Tours.

- une interface graphique évoluée masquant le langage de commande à l'utilisateur (basée sur X11), et permettant l'aide en ligne de manière navigationnelle.

- des échanges de données avec un tableur, augmentant ainsi les possibilités d'entrée de données.

- des pilotes graphiques pour des imprimantes souvent rencontrées sur les micro-ordinateurs. La mise au point de dessins s'opère ainsi sans quitter le bureau.

- la bibliothèque NAG[18] étant disponible sur nos serveurs, la possibilité d'appeler (plus ou moins) directement ses modules depuis des fonctions S.

4. Quelques considérations finales

Il peut apparaître que la maîtrise de S nécessite une logistique lourde, notamment en ressources humaines. A l'I.N.R.A., cela est indispensable pour supporter une volonté stratégique d'en faire l'un des produits utilisés par tous. Le choix initial de prendre entièrement en charge en interne les portages, les extensions, les mesures d'accompagnement, était rendu obligatoire par l'absence de distributeur en France.

Cela n'est plus vrai aujourd'hui[19], et un laboratoire aura sans doute grand intérêt à ne pas se préoccuper de ces aspects logistiques. S est très utilisé dans le monde universitaire et de la recherche, aux USA bien sûr, mais aussi en Allemagne, au Royaume Uni, en Australie, en Nouvelle-Zélande, au Japon. On peut espérer un rapide développement en France.

Enfin, de la même manière qu'il y a quelques années on publiait des algorithmes nouveaux en Fortran, Algol ou autre Pascal, l'existence de la bibliothèque publique *statlib*, la lisibilité et la puissance des fonctions écrites en langage S en font un vecteur privilégié de diffusion de résultats scientifiques.

C'est la notion de système d'accueil qui est ainsi mise en valeur :

- pour les développeurs qui trouvent un support d'intégration et de diffusion de leurs travaux

- pour les utilisateurs qui - de plus en plus - accèdent à un ensemble de progiciels différents (la boîte à outils) sans quitter un environnement de référence.

18 - NAG est une marque déposée du Numerical Algorithm Group.

19 - Une variante commerciale de S, *S-Plus*, est distribuée en France par la Société SIGMA+ de Toulouse.