Livre Blanc sur l'informatique en appui à la Recherche au CNRS

Pratiques, besoins, défis et recommandations

Mars 2014


cnrs
www.cnrs.fr

Livre blanc sur l'informatique en appui à la recherche au CNRS

Pratiques, besoins, défis et recommandations

Comité de Coordination et de pilotage de l'Informatique en Soutien à la recherche (CCIS) adossé au Comité d'Orientation pour le Calcul Intensif (COCIN)

Mars 2014

Composition du COCIN / CCIS au 1er janvier 2014

Représentants des Instituts du CNRS :

Michel Bidoit (INS2I) : Président du COCIN
Michel Daydé (INS2I) : Directeur du COCIN
Philippe Helluy (ISMI)
Pierre-Etienne Macchi (IN2P3)
Claude Pouchan (INC)
Denis Veynante (INSIS)
Stefano Bosi (INSHS)
Denis Girou (IDRIS)
Laurent Lellouch (INP)
Thierry Meinnel (INB)
Gudrun Bornette (INEE)
Jean-Pierre Vilotte (INSU)

Membres invités :

Vincent Breton (Directeur de l'IDGC)
Olivier Porte (DSI)

Experts :

Dominique Bascle (INC)
Maurice Libes (INSU)
Françoise Berthoud (INP)
Vincent Miele (INEE)

L'INFORMATIQUE SCIENTIFIQUE EN APPUI A LA RECHERCHE : Synthèse

1. Dans les laboratoires CNRS, l'informatique scientifique est déclinée en plusieurs spécialités diversement représentées au sein des instituts du CNRS : calcul scientifique ; développement logiciel ; bases de données ; interface homme-machine et interface web ; traitement, analyse et pérennisation des données scientifiques ; instrumentation et acquisition de données, architecture et infrastructure pour l'informatique scientifique. Ce contour n'inclut ni la recherche en informatique, ni l'informatique de gestion ou la gestion globale des systèmes d'information du CNRS.

2. L'informatique scientifique s'inscrit fortement dans le processus de recherche et est l'un des éléments fondamentaux de la compétitivité scientifique internationale. Notre enquête montre que les différentes communautés scientifiques associées aux instituts du CNRS se déclarent très impliquées dans tout ou partie des spécialités en informatique scientifique répertoriées ci-dessus et souligne son lien direct avec la qualité de la production scientifique dans un contexte concurrentiel mondialisé. Cette vision est partagée au sein des instituts qui ont une forte tradition en informatique scientifique (INSU, IN2P3...) mais également par d'autres instituts (INEE, INSHS...) pour lesquels des besoins plus récents en informatique scientifique s'expriment désormais et s'installent de manière pérenne.

3. Depuis la collecte, le stockage, le traitement et l'analyse des données jusqu'au développement de modèles numériques et aux problématiques d'accessibilité et de mise à disposition des données, l'informatique scientifique contribue à apporter des réponses à l'omniprésence de ces besoins liés aux données et allant jusqu'au "Big Data". Dans ce nouveau paysage, le manque de compétences et/ou de personnels sur les spécialités liées au traitement des données au sens large est un constat majeur, au vu des besoins qu'expriment la majorité des instituts.

4. Les personnels ingénieurs et techniciens en appui à la recherche sont au cœur de la mise en œuvre de l'informatique scientifique. Ils sont majoritairement et préférentiellement positionnés au plus près des problématiques scientifiques spécifiques à chaque institut. Par ailleurs, l'implication des chercheurs dans l'informatique scientifique est souvent significative et inversement corrélée au niveau de soutien en personnel technique : ce soutien faisant défaut dans nombre de laboratoires, il peut impacter l'évolution et le déroulé des projets de recherche.

5. De grandes différences selon les instituts sont observées dans les pratiques en informatique scientifique. Celles-ci se traduisent par des approches et des compétences hétérogènes. Par exemple, certains instituts affichent des compétences autour de la gestion et du traitement des masses de données qui seraient bénéfiques à l'ensemble des instituts. A l'inverse, d'autres instituts développent de fortes compétences autour du calcul scientifique. Pourtant, en dépit de ces différences, les besoins exprimés par l'ensemble des instituts sont assez semblables.

RECOMMANDATIONS

1. Le CNRS doit se positionner pour proposer une approche avec des objectifs ambitieux afin de faire face aux défis du monde de la recherche en ce qui concerne le référencement, le traitement et l'analyse des **données**. Cette impulsion stratégique conforterait les recommandations relatives aux besoins en matière de calcul et de données déjà identifiées dans le Livre Blanc du Calcul Intensif au CNRS. Face à la problématique des données, il s'agira notamment de **renforcer** et de **coordonner** d'une part **les moyens à allouer** et d'autre part la **diffusion des compétences** existantes à ce jour dans de nombreuses équipes.

2. Par sa vision stratégique, le CNRS devra favoriser le **partage et le transfert de compétences** en informatique scientifique entre instituts. Dans cet esprit, la **formation** doit être l'élément central et récurrent. Les instituts plus récemment concernés par l'informatique scientifique doivent être accompagnés pour organiser les formations nécessaires dans des domaines liés à leurs besoins, en impliquant les autres instituts. Les services de formation permanente devront par ailleurs enrichir leur offre de formation en Informatique Scientifique, en s'appuyant sur les compétences locales. Dans un souci d'optimisation des coûts, il devient impératif de tirer profit de collaborations inter-organismes pour faciliter l'accès à la formation au plus grand nombre.

3. Concernant les **personnels de soutien à la recherche**, le CNRS doit veiller à préserver une réponse adaptée aux besoins exprimés, particulièrement dans les secteurs où la demande émerge. Outre le **maintien ou la création de postes en Informatique Scientifique**, différentes approches complémentaires pourraient être explorées dans l'objectif de maintenir et de renforcer une expertise au plus près des chercheurs, et ce en facilitant les **échanges de compétences inter-instituts** : offre de **mobilité sur projets** de courte durée, structuration de pôles de compétences transversaux, etc. Une grande palette de possibilités est à explorer et à expérimenter. Par ailleurs, il est nécessaire de veiller à une collaboration forte et nécessaire entre toutes les composantes de l'informatique en appui ou en soutien à la recherche.

4. L'informatique scientifique doit être reconnue et considérée comme un véritable **pôle stratégique** au sein du CNRS. Dans ce but, il appartient au CNRS de conduire une réflexion autour de l'évolution des besoins et des métiers et des réponses à apporter. Il importera de mettre en œuvre des actions significatives permettant d'insuffler une réelle stratégie du CNRS transverse et adaptée aux besoins de chaque institut. Cette réflexion doit associer les divers acteurs (DSI, DIST, experts métiers) et les instituts dans le contexte complexe d'un paysage multi-organismes.

1. INTRODUCTION

Ce Livre Blanc a été réalisé en 2013 sur la base d'une enquête menée entre novembre 2012 et janvier 2013 permettant de collecter des données issues de l'ensemble des Instituts de Recherche du CNRS au travers de leurs unités.

L'étude sur l'évolution des besoins en informatique scientifique dans les laboratoires a été menée au sein du Comité de Coordination et de pilotage de l'Informatique en Soutien à la recherche (CCIS) adossé au Comité d'Orientation pour le Calcul Intensif (COCIN). Le CCIS est constitué du COCIN épaulé d'experts couvrant les divers domaines de l'informatique scientifique.

Ce comité a conduit une réflexion sur l'informatique scientifique en appui à la recherche définie de la façon suivante :

- calcul numérique,
- développement scientifique,
- Base de données,
- IHM et/ou interface web scientifique,
- traitements et/ou analyses de données,
- traitements et/ou analyses statistiques,
- traitements et/ou valorisations de données d'observation,
- pérennisation des données scientifiques,
- instrumentation,
- acquisition de données et/ou de signal,
- architecture et infrastructure pour le calcul haute performance,
- etc.

Il s'agit de dresser une cartographie des enjeux et des besoins qui découlent de l'Informatique scientifique. Le CCIS, adossé au COCIN qui coordonne les réflexions des 10 instituts du CNRS en matière de Calcul Intensif, tire de cette enquête un certain nombre de propositions qui seront soumises au collège de direction du CNRS.

Le taux de retour à cette enquête est de 25 % des unités, ce qui est satisfaisant. Cependant, le panel des unités ayant répondu est plus ou moins représentatif selon les instituts, ce qui introduit un biais non négligeable dans certains cas et nous a dissuadé d'insérer dans ce Livre Blanc une analyse très détaillée des résultats par institut.

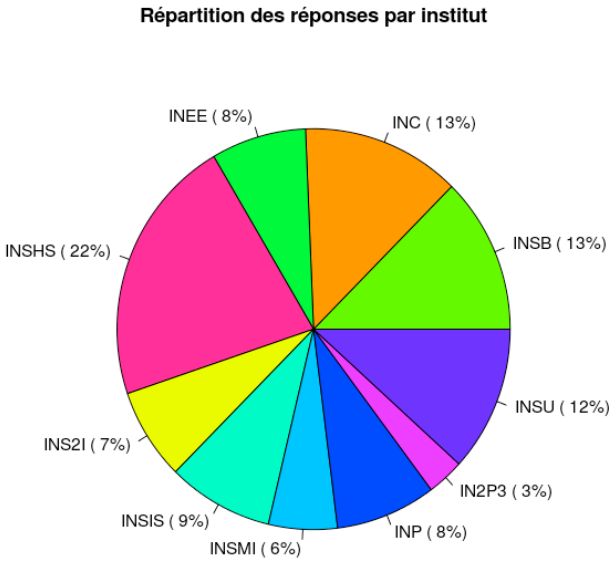
2. ANALYSE DES REPONSES A L'ENQUETE

2.1. Qui a répondu à l'enquête ?

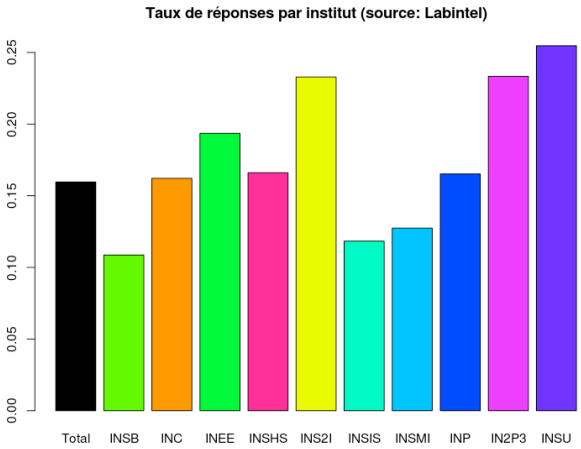
- 216 laboratoires représentant essentiellement des unités de recherche, plus quelques fédérations et autres structures ont répondu à cette enquête, soit près du quart des unités CNRS.
- La majorité des retours provient des directeurs de laboratoires (55 %) avec 1/3 des réponses apportées par les ingénieurs en informatique.

2.2. Les instituts ont-ils tous répondu de façon équivalente ?

Le volume des retours est très variable selon les instituts. En nombre absolu de réponses, INSHS est logiquement bien représenté, alors que en rapportant les chiffres au nombre d'unités dans chaque institut, il apparaît que ce sont INSU et INS2I qui ont le mieux répondu (voir graphique ci-dessous).



Cette figure donne le nombre de réponses par institut rapporté au nombre total de réponses à l'enquête. On constate le grand nombre de réponses pour l'INSHS.



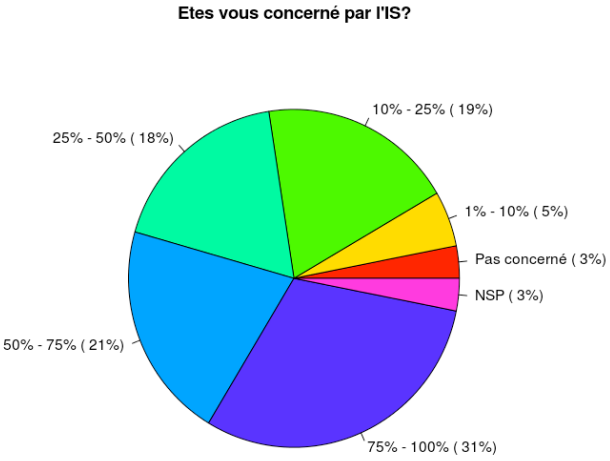
Rapporté au nombre de structures (unités de recherche, fédérations, GDS...) par institut, il apparaît que ce sont l'INSU et l'INS2I qui sont les mieux représentés (avec plus de 20 % des structures ayant répondu).

Remarque : ces résultats pourraient être affinés en fonction de la répartition laboratoires/autres structures selon les instituts.

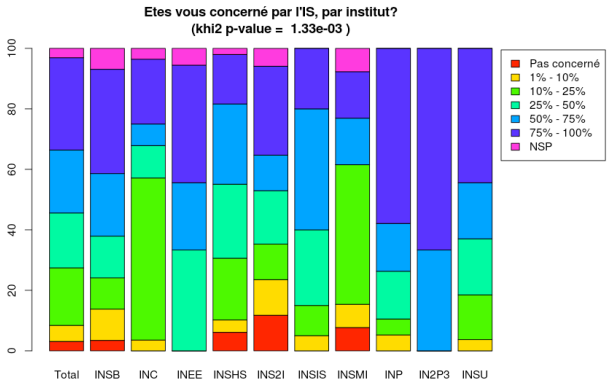
On notera qu'il n'y a pas de différences significatives du taux de réponse en fonction des délégations.

2.3. Est-ce que les sondés se sentent concernés par l'Informatique Scientifique (IS) ?

Plus de la moitié des unités qui ont répondu à l'enquête se sentent fortement concernées par l'IS, toutefois les résultats diffèrent significativement selon les instituts. Des instituts tels qu'IN2P3, INSU et INP se sentent largement concernés, alors que l'IS interpelle moins au sein d'INC et INSMI. Comme pour le taux de réponse à l'enquête, les résultats ne varient pas de façon significative selon les Délégations Régionales.



Plus de la moitié des entités qui ont répondu à l'enquête sont largement concernées par l'IS (plus de 50 %). Ce résultat doit permettre de mettre en perspective les réponses qui suivent.



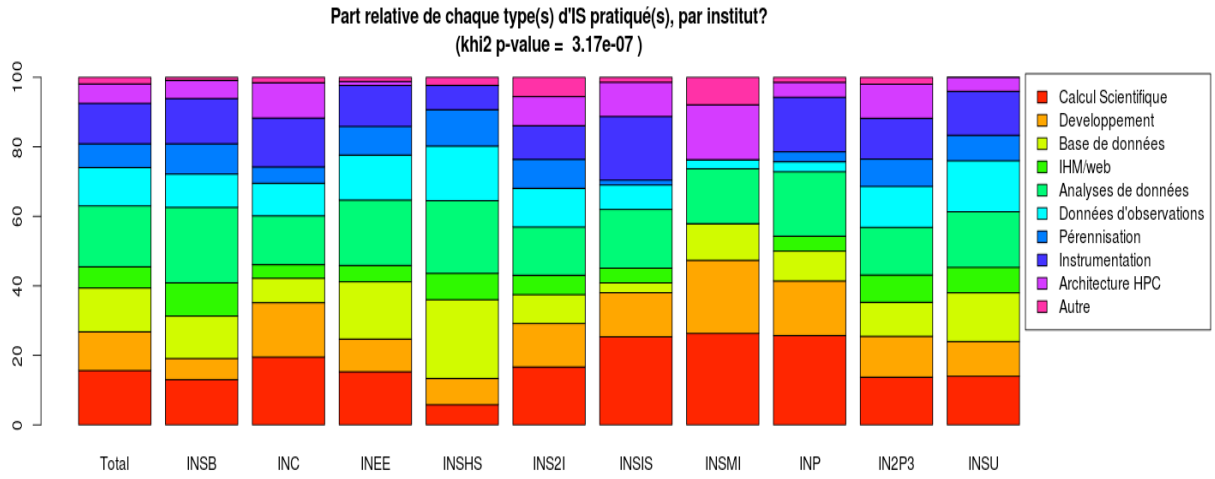
Les résultats diffèrent significativement d'un institut à l'autre : l'IN2P3, l'INSU et l'INP sont largement concernés tandis que pour les instituts de chimie et de mathématiques, l'informatique scientifique est moins présente dans leurs activités de recherche.

3. L'INFORMATIQUE SCIENTIFIQUE EN APPUI A LA RECHERCHE AU SEIN DES INSTITUTS DU CNRS : pratiques, enjeux, besoins, défis

Nous reprenons dans cette section point par point les retours aux diverses questions de l'enquête.

3.1. Pratiques liées à l'Informatique Scientifique

Il s'agit de mieux cerner quels types de techniques sont utilisés dans le cadre de l'informatique scientifique par les unités, les réponses pouvant être multiples. Nous proposons dans le graphique ci-dessous une analyse par institut.



Les écarts sont très significatifs entre les instituts qui exhibent donc des pratiques très diverses. Faisons quelques observations préliminaires :

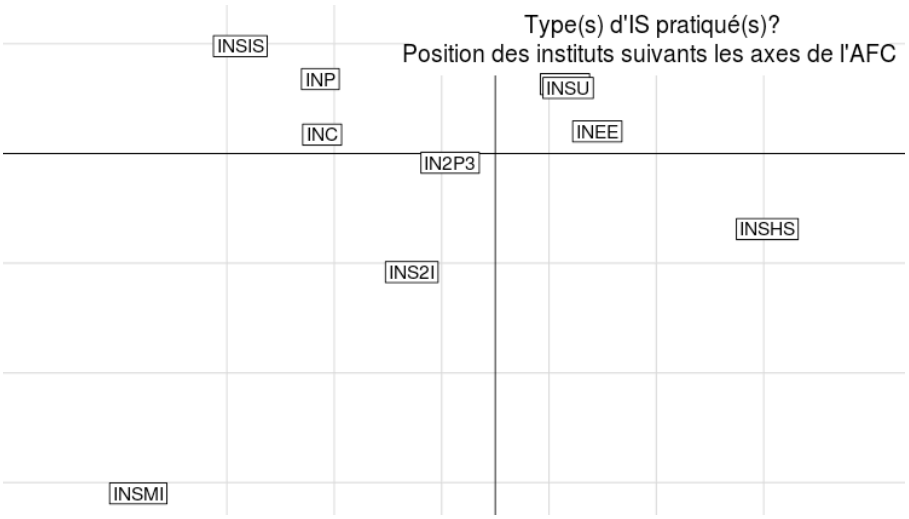
- Tous les laboratoires utilisent le développement scientifique, avec une importance relative plus forte pour l'INSMI.

- Le calcul scientifique est présent dans tous les instituts mais il reste assez minoritaire pour l'INSHS.

- Les techniques relatives à l'analyse de données (associée à des interfaces web) et à leur valorisation sont largement présentes dans tous les instituts. Il apparaît que les données sont souvent des données issues d'observations sauf pour les instituts de mathématiques et de physique.

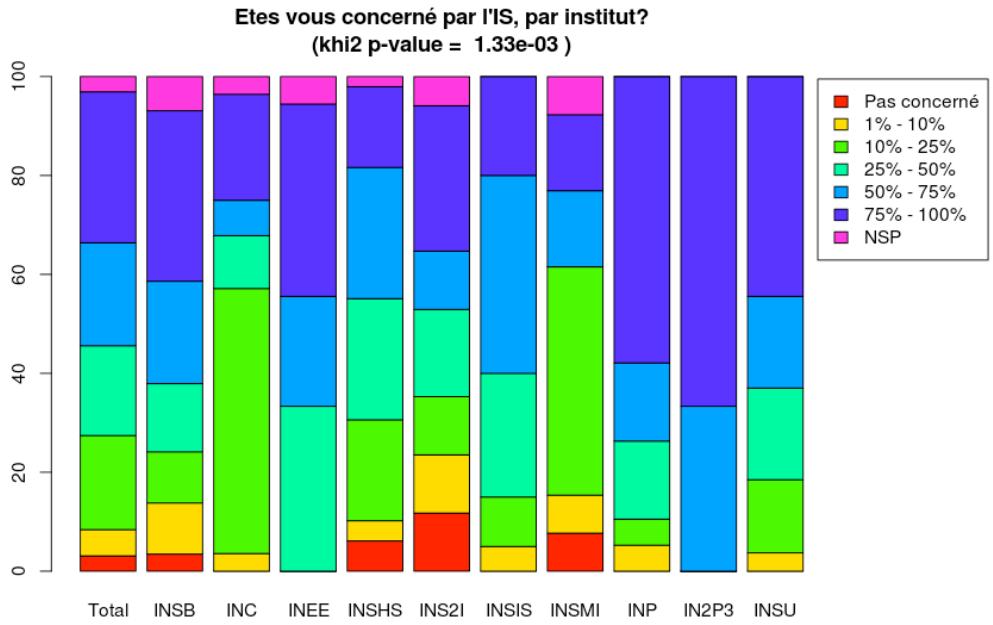
- L'INSMI se distingue aussi par l'absence d'informatique scientifique liée à l'instrumentation, qui est par contre significativement présente ailleurs. A l'inverse, l'INSMI se distingue par une grande importance donnée aux aspects matériels du calcul haute performance (ce qui peut paraître paradoxal parce que c'est, de tous les instituts CNRS, celui qui calcule le moins dans les centres nationaux mais est justifié de par l'importance des recherches liées au calcul haute performance au sein de cet institut).

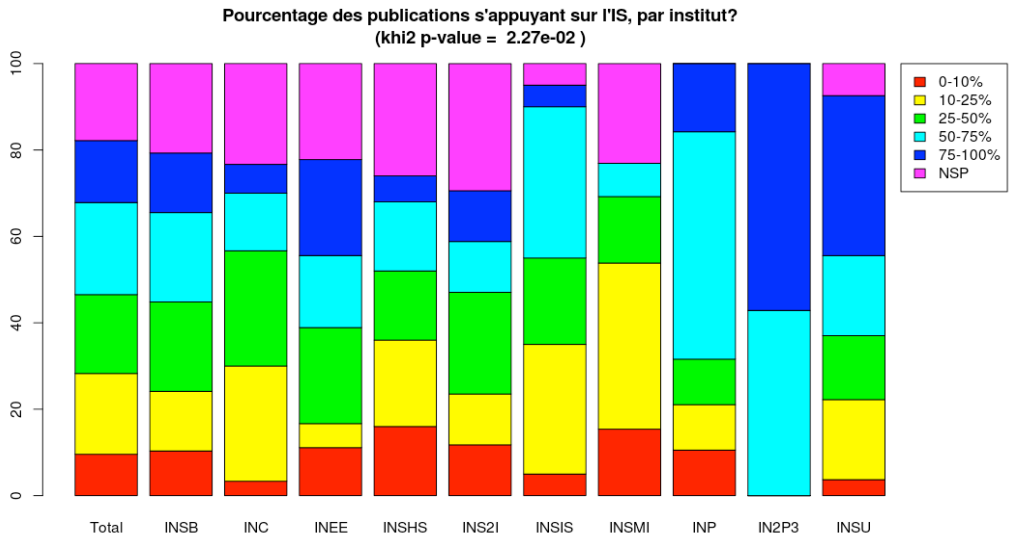
À partir de ces résultats, il est possible de positionner sur un plan les instituts à partir d'une analyse factorielle des correspondances. Cette figure permet de mettre en évidence les similarités/disparités entre instituts du point de vue des pratiques liées à l'informatique scientifique utilisées.



3.2. Impact de l'IS sur les publications scientifiques

Nous avons ensuite cherché à estimer l'importance de l'informatique scientifique dans les publications scientifiques. Ces résultats sont intéressants en les mettant en perspective avec les résultats de la question « Êtes-vous concerné par l'IS ? ». Ils sont représentés ci-dessous par institut.





On peut en tirer quelques remarques :

- Alors que d'une manière générale, les sondés savent répondre à la question de savoir si leur laboratoire est concerné par l'IS, il apparaît que plus de 20 % des sondés, dans les domaines de la biologie, de la chimie, de l'écologie, des sciences sociales, de l'informatique et des mathématiques ne sont pas en mesure de répondre à la question du pourcentage de publications s'appuyant sur l'IS. C'est un constat qui devrait interpeler. Est-ce que les apports de l'IS disparaissent entre le début des expériences/observations et les publications scientifiques qui en découlent ?

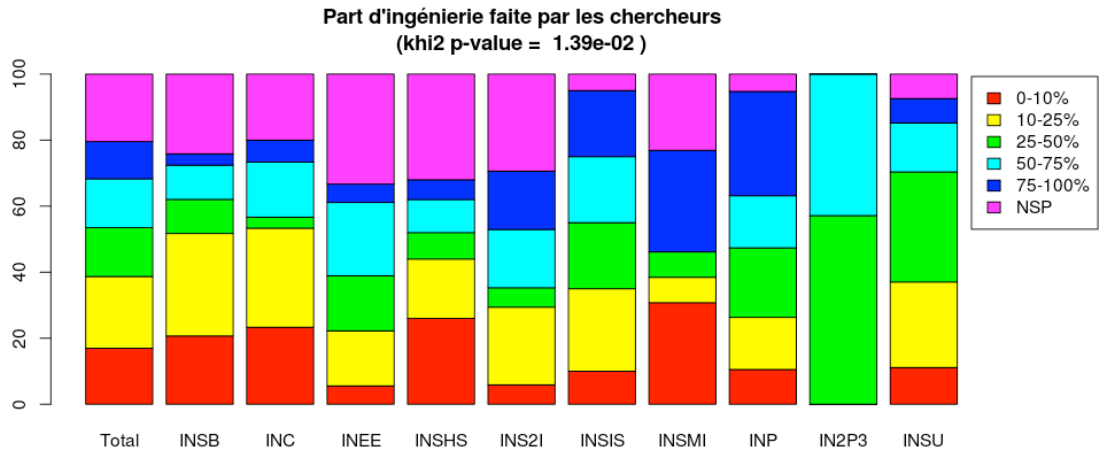
- À l'exception de l'INP, de l'INSU et de l'IN2P3 pour lesquels les techniques liées à l'IS sont bien implantées depuis plusieurs décennies, il y a un écart sensible entre les réponses aux deux questions !

3.3. Part d'informatique scientifique effectuée par les chercheurs eux-mêmes

Les résultats sont reportés par institut dans le graphique ci-dessous. On peut en tirer deux remarques :

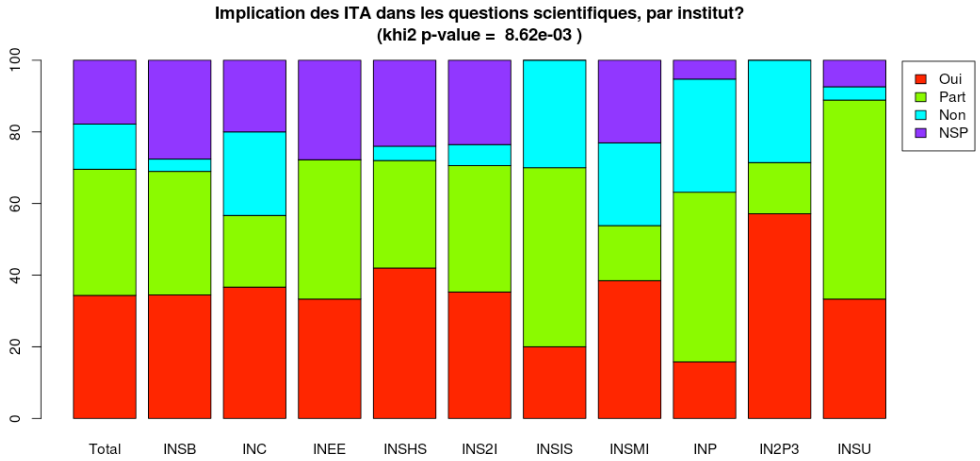
- Comme à la question précédente, nous notons une relative méconnaissance des liens entre chercheurs et IS, notamment pour les instituts de biologie, chimie, sciences humaines et sociales et mathématiques avec un fort pourcentage de réponses « NSP ».

- Il ressort que pour l'IN2P3, l'INSU, l'INSMI, l'INSIS et l'INEE, plus de 50 % de l'IS (en moyenne) est faite par les chercheurs eux-mêmes, contrairement à l'INC et à l'INSB où c'est la tendance inverse qui est observée.



3.4. Implication des ITA dans les questions scientifiques

Là encore, on constate de grosses différences culturelles selon les instituts avec une implication des ITA dans les questions scientifiques moins marquée pour l'INP et l'INSIS, voire pour l'INC et plus importante pour des instituts tels INSU et IN2P3.



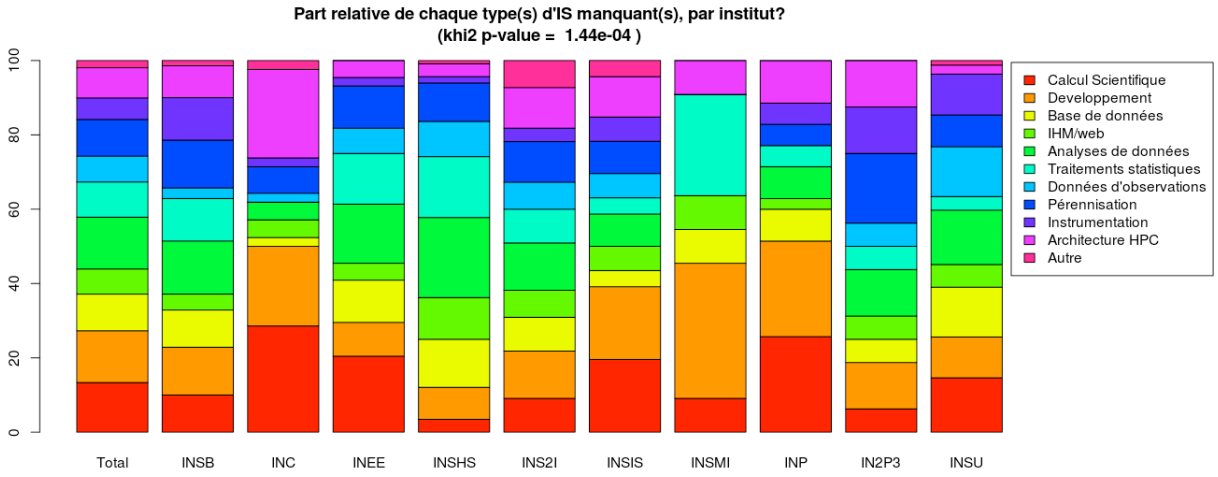
Enfin, il n'y a pas de différences significatives entre les instituts concernant le niveau de formation des IT dans le domaine scientifique concerné : en moyenne, 1/3 des sondés indiquent exiger un niveau de maîtrise ou d'expertise, 1/3 un niveau de base, 10 % ne demandent aucune connaissance particulière et les autres ne se prononcent pas.

3.5. Techniques de l'Informatique Scientifique faisant le plus défaut et impact sur l'organisation de la recherche

Les sondés devaient citer les techniques de l'informatique scientifique leur faisant le plus défaut. Les résultats synthétiques sont reportés dans le graphique ci-dessous par institut.

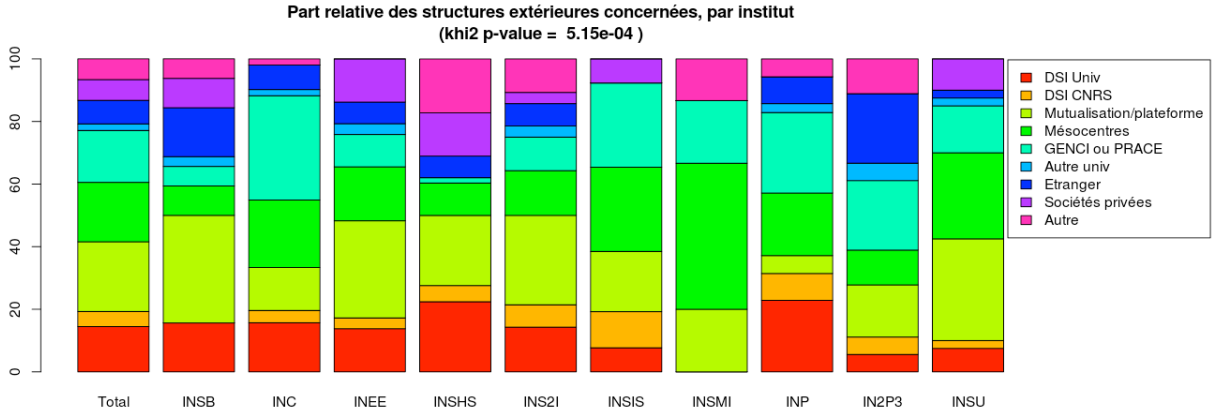
Les laboratoires de l'INP, de l'INSMI, de l'INSIS et de l'INC, voire l'INEE semblent plutôt déficitaires en techniques de calcul scientifique (et architecture HPC associée) et en techniques de développement. Pour l'INSU, l'IN2P3, l'INS2I, l'INSHS et l'INSB l'ensemble

des techniques font défaut de façon plus ou moins équilibrée. On constate aussi des besoins récurrents autour de l'analyse/traitements de données et aussi des IHM et du Web (sans doute liés aussi aux données) sur l'ensemble des instituts excepté INC où ils restent faibles. Les besoins liés à de l'instrumentation sont concentrés sans surprise sur INSB, INSIS, INP, IN2P3 et INSU.



Les sondés étaient questionnés sur les raisons principales de ces manques. Sans surprise, le déficit de moyens humains est ressorti dans près de 70 % des cas. Les autres raisons invoquées sont des problèmes d'infrastructures ou de matériels insuffisants (plus de 40 % des cas) ou de compétence (25 % des cas).

Pour faire face à ces difficultés, les laboratoires sont amenés à faire appel à des structures externes au CNRS, selon des stratégies significativement différentes selon les instituts :



On peut noter :

- Le recours pour une part importante aux services (DSI) des universités dans près de 15 % des cas en moyenne sauf pour l'INSMI.
- Une sollicitation importante vers les méso-centres, les plateformes mutualisées et les centres nationaux avec des profils variables selon les instituts.

Une utilisation non négligeable de structures localisées à l'étranger, notamment pour l'IN2P3 et l'INSB et dans une moindre mesure l'INP, l'INC, l'INEE, l'INSHS et l'INS2I ; par contre cette sollicitation est faible pour l'INSU et nulle pour l'INSMI.

4. CONCLUSION

Cette enquête n'a fait que confirmer l'importance de l'Informatique Scientifique en appui à la recherche pour la compétitivité des recherches menées au sein du CNRS. Nous en cernons mieux la physionomie et avons pu mesurer l'ampleur des besoins qui concernent très largement la valorisation des données ainsi que leur accessibilité au travers du Web, le développement logiciel, le calcul scientifique ainsi que des besoins plus spécifiques autour de l'instrumentation et l'imagerie.

Le rôle des ITA impliqués dans l'Informatique Scientifique est fondamental dans les avancées de la recherche. Préserver la compétitivité du CNRS passe par une réponse aux besoins exprimés en termes d'Informatique Scientifique ainsi que par une meilleure synergie entre les instituts. Il faut aussi soutenir des actions de formation en phase avec les besoins favorisant l'implication des ITA ainsi que porter une attention soutenue à l'évolution des carrières et des métiers.

Remerciements

Le COCIN / CCIS remercie Catherine Blanc (IRIT) pour la mise en forme du document final.