numéro 19

1024

BULLETIN
de la société informatique
de France

Avril *2022*

COMITÉ DE RÉDACTION

SOCIÉTÉ INFORMATIQUE DE FRANCE
Institut Henri Poincaré, 11 rue Pierre et Marie Curie, 75231 Paris Cedex 05

Prix public : 32 € (adhérents SIF : −30%)
Directeur de la publication : Yves Bertrand ISSN : 2270-1419

Couverture : d'après une maquette réalisée par Lollygraph.com.

SOMMAIRE DU Nº 19

PRIX ET DISTINCTIONS

PROJETS SCIENTIFIQUES

SCIENCE

RÉCRÉATION

Le hasard du calendrier fait que ces lignes sont écrites au moment même où le bruit assourdissant des bottes qui marchent sur l'Ukraine n'a pour écho que la difficulté des démocraties à les faire taire. Formons le voeu qu'à l'heure où vous nous lirez, l'Ukraine sera sortie du conflit en nation libre et indépendante.

Éditorial

Chères adhérentes, chers adhérents, chères lectrices, chers lecteurs,

À l'occasion des 10 ans d'existence de la SIF, nous vous offrons un exemplaire de la bande-dessinée des « Décodeuses du numérique », réalisée par le CNRS en collaboration avec la SIF (pour les fiches pédagogiques qui l'accompagnent, notamment), afin de lutter contre les stérotypes qui empêchent les jeunes filles et femmes de s'orienter vers les métiers du numérique. Nous vous l'avions déjà présentée lors du dernier numéro [1] mais recevoir cette BD vous permettra, nous l'espérons, de renforcer cette action et d'en parler autour de vous.

Par ailleurs, le comité de rédaction est à nouveau fier — pour plusieurs raisons — de vous faire parvenir le bulletin numéro 19 (○○○●○○●● en binaire pour ceux qui n'auraient pas encore fait le rapprochement) qui comporte un certain nombre de nouveautés que je vous propose de passer en revue maintenant.

La première — qui aurait pu passer inaperçue mais qui a été sérieusement discutée au sein du comité de rédaction de 1024 et du conseil d'administration de la SIF — réside dans la modification de la licence de réutilisation et de distribution des articles en passant de Creative Commons Attribution Pas de modification (CC BY-ND) à Creative Commons Attribution Pas d'utilisation commerciale Pas de modification (CC BY-NC-ND). Cette décision se justifie par le souhait de protéger les auteurs des articles de 1024. Cette licence n'interdit pas l'utilisation commerciale des articles

1. https://doi.org/10.48556/SIF.1024.18.

DOI:10.48556/SIF.1024.19.3

mais l'approbation des auteurs est alors nécessaire. Sauf demande expresse de l'auteur d'appliquer un autre type de licence, tous les articles publiés dans ce numéro et les suivants, jusqu'à nouvel ordre, le seront sous CC BY-NC-ND.

Vous avez pu découvrir la deuxième nouveauté depuis un an sur le site du bulletin et sur la version papier du précédent numéro qui consiste en un petit changement mais réalise un pas de géant pour le référencement de 1024. Eh oui ! Notre société savante s'est offert le luxe d'attribuer un *Digital Object Identifier*[2] (DOI) à chaque article et chaque bulletin de 1024. Pour cela, nous avons dû modifier la présentation numérique du bulletin en créant une page spécifique appelée « page d'atterrissage » (ou *landing page*) où vous pouvez, par exemple, télécharger la référence BibTEX de votre article. La raison d'être d'une telle démarche réside dans la volonté de valoriser et d'améliorer le référencement des articles publiés dans notre bulletin. Nous ne sommes pas les premiers et d'autres sociétés savantes l'ont fait avant nous (SMF, SFP...). La possibilité d'ajouter un nouvel article à votre bibliothèque HAL[3] simplement en renseignant cet identifiant (un ensemble de méta-données — titre, auteurs, numéros de pages, volume... — est associé à chaque DOI) était une autre de nos motivations. Malheureusement, pour des raisons techniques et de normalisation, ce service n'a pas été rendu possible en 2021 malgré la création de l'ensemble des DOI de *tous* les articles publiés dans 1024 depuis son origine. C'est pour cette raison qu'en janvier 2022 nous avons changé d'agence d'enregistrement de ces identifiants en adhérant à Crossref[4]. Le référencement et la visibilité de 1024 est en marche... et ce n'est pas fini !

Après cette période morose, nous avons voulu égayer notre bulletin en illustrant de manière décalée certains articles (bien entendu, avec l'accord des auteurs). Le plus difficile pour y arriver a évidemment été de trouver la perle rare souhaitant s'atteler à cette tâche et nous l'avons dénichée parmi notre communauté de chercheurs et d'enseignants-chercheurs : Frédéric Havet, qui nous fait profiter de son talent de dessinateur et d'humoriste. Une motivation de plus pour feuilleter notre bulletin.

176 pages !

En janvier 2022, jamais je n'aurais pensé le voir aussi volumineux et, à nouveau, 1024 est rempli d'articles très intéressants. Je tiens à remercier les auteurs pour nous avoir fait confiance en proposant leur texte et les personnes qui les ont sollicités. Par ailleurs, je tiens également à remercier la personne qui relie de manière très approfondie tous les textes de 1024 pour y apporter des corrections orthographiques, typographiques, stylistiques... Bien évidemment, si vous voyez une erreur dans ces deux dernières phrases, sachez que Patrice Naudin ne les as pas relues...

2. https://fr.wikipedia.org/wiki/Digital_Object_Identifier.
3. https://hal.archives-ouvertes.fr.
4. https://www.crossref.org.

Ce numéro ne contient pas explicitement de dossier thématique comme les précédents mais tente de renforcer les rubriques scientifiques en vous présentant les résultats de deux projets de recherche ANR passés : le premier sur les nouvelles techniques de manipulation de contenus numériques sur une surface plane et le second sur la gestion des connaissances et la production de logiciels en pharmacogénomique. De plus, nous avons souhaité donner la parole, ou plutôt la plume, aux collègues récemment habilités à diriger des recherches. Découvrez donc les contradictions de l'apprentissage soit-disant non-supervisé, la récurrence nœthérienne avec un point de vue très formel ou encore une vision complémentaire au thème « Reproductibilité » abordé dans le précédent numéro sur la crédibilité des résultats de simulations numériques.

Les autres rubriques ne sont pas en reste avec les résumés de nos jeunes docteurs primés par le prix de thèse Gilles Kahn ou encore un article d'actualités sur la réduction, à toutes les étapes de conception, de l'impact environemental d'un logiciel...

Comment la science a-t-elle été perçue et vécue en temps de crise ? Voilà à quoi s'intéresse la journée Sciences et média de 2022 et nous vous offrons le discours inaugural de cette journée, prononcé par des journalistes scientifiques.

Retrouvez aussi la rubrique que nous souhaiterions la plus petite possible, rédigée par des personnes souhaitant dire un dernier mot à leur ami ou collègue regretté.

Enfin, lisez, réfléchissez et envoyez vos réponses à la fameuse et tant attendue rubrique de récréation qui s'intéresse cette fois-ci aux puissances de 2... Tiens, cela me rappelle quelque chose !

Il est temps de vous souhaiter une très agréable lecture en vous rappelant que le bulletin « 1024 » vit et évolue pour vous et grâce à vous, alors n'hésitez pas à nous proposer vos idées, expériences ou réflexions et à encourager votre entourage à faire de même. Plus tôt nous recevrons votre texte, plus il aura de chances d'être illustré.

DENIS PALLEZ
Rédacteur en chef
1024@societe-informatique-de-france.fr

Le mot du président

Yves Bertrand [1]

Chères adhérentes, chers adhérents,

Le congrès des 10 ans de la SIF arrive à grands pas. Nous espérons vous y voir nombreux. Mais l'élection présidentielle survient plus vite encore, charriant surenchères verbales et propositions déraisonnables. C'est dans ce contexte que je vous propose de vous entretenir d'un sujet qui défraya la chronique plusieurs semaines en ce début d'année 2022, via d'innombrables médias régionaux et nationaux, pour questionner de façon plus large le rôle d'une association telle que la nôtre. Le sujet, donc : *les mathématiques dans la réforme du lycée général*. Le constat des mathématiciens est triple :

— une part significative des élèves abandonnent les mathématiques dès la première et près d'un élève sur deux n'étudie plus les mathématiques en terminale ;

— cet abandon est très marqué pour les filles ;

— en moyenne chaque élève apprend bien moins de mathématiques qu'avant la réforme.

En quoi sommes-nous concernés en tant qu'informaticiens ?

Nous sommes concernés parce que la spécialité NSI est encore étique en termes d'effectifs (par rapport aux spécialités mathématiques, physique-chimie et SVT), même si cela est compréhensible pour une spécialité qui n'avait pas de réel équivalent avant la réforme.

Nous sommes concernés parce que la question abordée est, plus largement, celle de la formation scientifique des 90 % de lycéens non prédestinés à l'entrée en classe préparatoire scientifique. Ils doivent tous bénéficier d'un bagage scientifique solide

1. Président de la Société informatique de France, professeur des universités, université de Poitiers.

DOI:10.48556/SIF.1024.19.7

pour être mieux armés dans leur vie quotidienne et avoir une attitude scientifique face aux *fake-news* dont les réseaux sociaux nous inondent.

Nous sommes concernés en tant que scientifiques inquiets des conséquences potentielles d'une réforme qui — en l'état — risque de ne pas former assez de cadres capables de relever les défis sociétaux et économiques du XXI^{e} siècle, comme l'illustre le communiqué de la SIF, page 11 de ce bulletin, appelant à « Faire de l'enseignement scientifique une grande cause nationale ».

C'est pourquoi, avec l'appui de membres du conseil scientifique et du conseil d'administration de la SIF, nous avons communiqué d'abord avec les mathématiciens [2], puis peu après avec une vingtaine d'associations ou sociétés savantes scientifiques [3] sur certains écueils que présente la réforme du lycée général. L'impact de la communication fut tel qu'un comité *ad hoc* a été mis en place par le MENJS, avec pour consigne de fournir au ministre, pour le 15 mars, des éléments d'amendement de la réforme applicables à la rentrée 2022. Au moment où vous lisez ces lignes, ce comité aura donc remis ses conclusions.

Ces circonstances exceptionnelles posent avec acuité la question du rôle même d'une société savante telle que la SIF. Je vous soumets trois axes de réflexion :

— elle doit s'inscrire dans un temps propice à une réflexion de fond, et non dans un temps contraint par une échéance électorale, fût-elle présidentielle ;

— elle doit promouvoir des actions structurantes, plus que des ajustements marginaux qui exonèreraient d'une réelle prise en compte des problèmes soulevés ;

— elle doit empreindre son action et ses propositions de valeurs humanistes universelles, indépendamment de toute contingence politicienne.

En particulier, en accord avec le collectif scientifique engagé dans les alertes sur les impacts de la réforme du lycée, notre action vise à mettre en place un groupe de travail sur une période enjambant les échéances électorales, dont la réflexion puisse questionner puis améliorer la structure même de la réforme à l'horizon 2023, prioritairement sur les points suivants :

— le nécessaire équilibre entre sciences et humanités, actuellement en défaveur des sciences ;

— la possibilité d'élargir à trois spécialités (au lieu de deux) la formation des lycéens en terminale ;

— des mesures garantissant une meilleure intégration des filles aux parcours scientifiques.

---

2. https://www.societe-informatique-de-france.fr/wp-content/uploads/2022/02/2022-02-07-communiquesynthese_math_filles.pdf.

3. https://www.societe-informatique-de-france.fr/wp-content/uploads/2022/02/2022-02-18-Synthese_Sciences.pdf.

Plus généralement, notre société savante, qui *sait*, justement, ce qu'*est* et ce que *peut* l'informatique, sera plus dans son rôle en défendant le logiciel libre plus que le logiciel propriétaire, les publications ouvertes à tous plus que la dépendance aux éditeurs oligarques, l'autonomie numérique plus que la dépendance aux GAFAM, et en amont, une école de l'égalité des chances, des territoires et des genres plus qu'une école des élites.

Ces objectifs ne sont bien entendu pas neutres. C'est pour cela que je les soumets à votre jugement, à votre éventuelle désapprobation, à débat — par exemple lors de notre congrès des 10 ans de la SIF ou dans les pages de 1024 — parce que c'est dans cet état d'esprit que je tente d'animer notre communauté.

Faire de l'enseignement des sciences une grande cause nationale

SIF

Nous vivons aujourd'hui une révolution scientifique et technique sans précédent dans notre histoire. Sans précédent parce qu'elle touche tous les aspects de notre existence : notre manière d'apprendre, de nous soigner, de produire des biens et des services, d'échanger avec nos proches, d'organiser la vie de la cité, de produire et distribuer des œuvres d'art... Cette révolution s'appuie sur l'ensemble des sciences et des techniques : l'informatique bien entendu, mais aussi les mathématiques, la physique, la chimie, les sciences de la vie et de la Terre...

Cette révolution ne fait que rendre plus inquiétants les piteux résultats de la France, année après année, dans les enquêtes Times, Pisa... Si la France cesse d'être une nation scientifique et technique de premier plan, c'est ailleurs (en Asie, en Amérique) que se dessineront les contours de cette révolution et que se fabriqueront les objets matériels et logiciels que nous utiliserons et c'est ailleurs qu'ira la richesse incommensurable qu'engendre cette industrie. Si toutes les citoyennes et tous les citoyens n'ont pas acquis à l'école quelques éléments de culture scientifique, ils utiliseront ces objets de manière passive, sans en comprendre les enjeux.

Une révolution de moins grande ampleur (la conquête spatiale) avait suscité, à la fin des années 1950, une transformation complète de l'enseignement des sciences dans les pays « de l'ouest », après leur prise de conscience, lors du lancement du premier satellite Spoutnik, le 4 octobre 1957, de leur retard par rapport aux pays « de l'est ». C'est d'une prise de conscience similaire dont nous avons besoin aujourd'hui.

Malheureusement, malgré un certain nombre d'aspects positifs, la récente réforme du lycée et du baccalauréat général ne permet pas de renforcer cette prise de conscience de l'importance des sciences et des techniques dans la formation des

DOI:10.48556/SIF.1024.19.11

femmes et des hommes du XXI^{e} siècle : l'emploi du temps d'un ou d'une élève de première ou de terminale est composé (hors éducation physique et sportive) d'un enseignement électif de 12 heures (les spécialités) et d'un enseignement de tronc commun obligatoire de 14 heures réparties en 12 heures consacrées aux humanités (français, philosophie, langues, histoire et géographie, enseignement moral et civique) et 2 heures consacrées aux sciences. Il est gravement anachronique, au XXI^{e} siècle, d'imaginer que l'enseignement reçu par une lycéenne ou un lycéen soit centré uniquement sur les humanités. Cela a de nombreuses conséquences néfastes : les étudiantes et étudiants qui s'orientent après le bac vers des études littéraires (droit, lettres, langues, philosophie, histoire...) auront un niveau scientifique comparable à celui d'un ou d'une élève de seconde, alors que le droit, l'éthique, les arts... sont transformés par la révolution scientifique et technique que nous vivons. Pire : les étudiantes et étudiants qui se destinent aux métiers de professeur des écoles auront également un très faible niveau scientifique et technique, risquant de perpétuer cette culture de l'ignorance scientifique.

L'avenir d'une nation se construit avant tout en dispensant à sa jeunesse une formation correspondant aux modifications majeures qui la traversent. Sachons lui proposer une formation à la hauteur des enjeux économiques et sociétaux du XXI^{e} siècle, qui nécessitent un juste équilibre entre humanités et sciences, indispensables les unes et les autres. Si nous voulons demeurer un pays hautement industrialisé et une société où les citoyens sont à l'aise avec les innovations scientifiques et technologiques il est indispensable que nous fassions de l'enseignement des sciences la grande cause du quinquennat à venir et que nous commencions par rééquilibrer le tronc commun du lycée d'enseignement général : 7 heures pour les sciences et 7 heures pour les humanités.

Le 24 janvier 2022

Avancement de grade des carrières universitaires

SPECIF Campus [1] et SIF

Le 27 janvier 2022, le conseil d'administration de la SIF a voté en faveur de la signature de la pétition proposée par la Commission permanente du Conseil national des universités à propos de l'avancement de grade des carrières universitaires [2] et propose un communiqué d'explication de ce vote. L'association SPECIF Campus a souhaité être associée à ce texte.

La Commission permanente du Conseil national des universités (CP-CNU) appelle à signer une pétition « pour l'équilibre entre le local et le national » pour l'avancement de grade des carrières universitaires[2]. Avec la LRU, les universités ont certes gagné en autonomie, et donc en moyens de développer plus efficacement leur propre politique de formation et de recherche, notamment en termes de pilotage des carrières de leurs chercheurs et enseignants-chercheurs.

Mais au-delà de la question spécifique de l'avancement de grade, lorsque cette autonomie conduit à la gestion exclusivement locale des carrières, sans régulation nationale opérée par le CNU, elle fait courir le risque du localisme, voire du clientélisme. A contrario et plus généralement encore, l'équilibre des missions, prérogatives et moyens entre les établissements locaux et les structures et organismes nationaux nous semble être la condition sine qua non de l'existence d'un pilotage stratégique national pertinent dont notre pays a besoin.

1. Société professionnelle des enseignants et chercheurs en informatique de France, https://www.specifcampus.fr.

2. https://www.wesign.it/fr/education/carrieres-universitaires--pour-lequilibre-entre-le-local-et-le-national.

DOI:10.48556/SIF.1024.19.13

Transformer les structures nationales en coquilles vides et les organismes nationaux en agences de moyens constitue le chemin le plus sûr pour abandonner ce pilotage stratégique. Il est pourtant indispensable au niveau de chaque discipline, et en particulier à celui de l'informatique, tant elle irrigue l'ensemble des champs scientifiques, économiques et éthiques de notre société moderne.

Signer la pétition de la CP-CNU apparaît à SPECIF Campus et à la Société informatique de France comme l'un des moyens de ne pas s'engager sur un tel chemin. C'est pourquoi nos conseils d'administration respectifs ont voté le fait de signer la pétition de la CP-CNU au nom de nos associations. SPECIF Campus et la SIF soutiennent ainsi, formellement, une démarche visant à garantir une gestion équilibrée des carrières universitaires entre le local et le national.

Systèmes de saisie de texte pour les personnes avec une déficience motrice : comment les systèmes de prédiction linguistique contribuent-ils à améliorer les performances de saisie ?

Jean-Yves Antoine [1] et Mathieu Raynal [2]

Introduction

Avec l'utilisation croissante de dispositifs numériques (ordinateurs de bureau, portables, smartphones, tablettes, etc.), la saisie numérique de texte ou de données est devenue une tâche dominante dans notre quotidien, que ce soit pour travailler (rédaction de documents) ou pour communiquer (mails, réseaux sociaux, etc.). Pour cela, nous avons pris l'habitude d'utiliser un clavier. Lorsque le dispositif que nous utilisons ne dispose pas de clavier physique, ce dernier est généralement remplacé par un clavier dit logiciel : c'est-à-dire un clavier affiché à l'écran avec lequel l'utilisateur peut interagir soit au travers d'un écran tactile soit au moyen d'un dispositif de pointage.

Pour les personnes ayant une déficience motrice des membres supérieurs, le seul moyen de saisir sur un appareil informatique est le clavier logiciel qu'ils peuvent manipuler avec un dispositif de pointage adapté à leurs capacités motrices (joystick, système de suivi du regard, etc.). Pour les personnes ayant une très faible motricité, il

1. Université de Tours, Jean-Yves.Antoine@univ-tours.fr.
2. Université de Toulouse — IRIT, mathieu.raynal@irit.fr.

DOI:10.48556/SIF.1024.19.15

est également possible de remplacer le dispositif de pointage par un système de défilement automatique où le curseur se déplace de touche en touche et que l'utilisateur peut arrêter sur le caractère souhaité en utilisant un contacteur.

Dans ces différentes situations de handicap, la saisie sur clavier logiciel se révèle bien plus lente que ce qu'elle peut être avec un clavier physique. Cependant, pour augmenter la vitesse de saisie, les claviers logiciels ont généralement recours à des systèmes de prédiction linguistique pour faciliter l'accès aux caractères suivants ou proposer les mots à venir. Dans cet article, nous dressons un état des lieux des travaux existants sur l'intégration des systèmes linguistiques dans les claviers logiciels et la manière dont les résultats issus de ces systèmes sont exploités par le clavier logiciel.

Couplage d'un système de prédiction à un système de saisie de texte

Utilisation de la prédiction de mots

La technique la plus couramment utilisée pour accélérer la saisie de texte est de présenter une liste des mots les plus probables que l'utilisateur est en train de saisir. Si le mot que l'utilisateur souhaite saisir se trouve dans la liste, il peut alors le sélectionner et ainsi lui éviter de taper la fin du mot au clavier. Si le mot désiré n'est pas dans la liste, l'utilisateur continue d'entrer les caractères au clavier et le système met à jour la liste des mots en fonction du préfixe saisi. L'objectif est de prédire le mot que l'utilisateur souhaite saisir. Comme annoncé dans [9], les listes de prédiction de mots visent à accélérer la vitesse de saisie de texte en réduisant le nombre de clics nécessaire pour saisir le même texte sur un clavier ordinaire. Cela peut également réduire le nombre d'erreurs que l'utilisateur peut commettre au cours de sa saisie. Les listes de mots peuvent être utilisées avec tout type de système de saisie de texte et quelle que soit la modalité d'interaction utilisée. Par exemple, le système Sibylle [23] (cf. figure 1) est un système basé sur le défilement d'un curseur qui se déplace automatiquement d'une touche à une autre. La liste de mots proposés est intégrée au système de manière à ce que le curseur passe également par chaque touche liée aux différents mots. Le système HandiAS [13] est quand à lui un clavier logiciel augmenté d'une liste de mots prédits. Cette liste est positionnée de manière fixe à gauche du clavier. Le clavier PoBox [12] a lui été conçu pour une utilisation sur appareil mobile. Le clavier est affiché sur l'écran tactile de l'appareil et l'utilisateur peut utiliser ce clavier en utilisant un stylet. Ce clavier a comme originalité de proposer la liste de mots toujours au plus près de la dernière touche saisie par l'utilisateur. Cette technique a l'avantage de réduire la distance entre la saisie de l'utilisateur et la liste de mots, mais présente l'inconvénient de cacher une partie du clavier lorsque la liste est affichée.

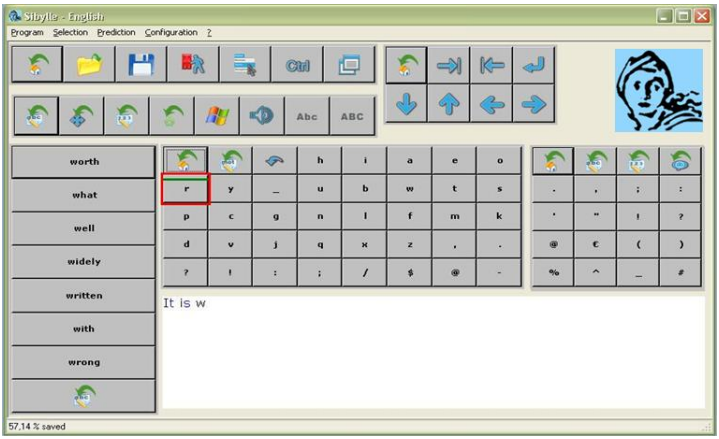


FIGURE 1. Sibylle avec le pavé central qui permet de saisir les caractères alphabétiques, et la liste de mots prédits à gauche. Le curseur, à droite de *what*, se déplace automatiquement de touche en touche.

Dans tous ces systèmes, l'utilisateur ne peut sélectionner dans la liste qu'un mot entier. Par conséquent, tant que le mot désiré n'est pas dans la liste, l'utilisateur est obligé de continuer de saisir le préfixe de son mot sur le clavier pour que le système de prédiction mette à jour la liste des mots et ce, jusqu'à ce que le mot souhaité soit affiché dans la liste. Dans ce cas de figure, l'utilisateur perd du temps à saisir son mot, du fait de la recherche visuelle du mot dans la liste qui n'abouti pas à la saisie du mot dans la liste. Pour remédier à ce problème, le système WordTree [4] propose une liste de mots où l'utilisateur peut ne sélectionner qu'une partie du mot qu'il souhaite saisir. Cette technique permet ainsi à l'utilisateur de sélectionner une chaîne de caractères dans la liste sans que celle-ci soit un mot complet. Ainsi cela réduit le nombre d'actions à réaliser pour saisir le mot.

Utilisation de la prédiction de caractères

Les systèmes de saisie utilisant les résultats des algorithmes de prédiction de caractères sont moins répandus que ceux basés sur la prédiction de mots. Ils sont surtout utilisés dans le cas des claviers logiciels utilisant le défilement automatique du curseur. Dans ce cas précis, le curseur se déplace de touche en touche à intervalle régulier, généralement de gauche à droite et de bas en haut sur le clavier. Après chaque caractère saisi, le curseur est toujours repositionné dans le coin supérieur gauche du clavier. L'utilisateur doit alors attendre que le curseur arrive sur le caractère souhaité pour le valider au moyen d'un contacteur. L'algorithme de prédiction de caractères

est donc utilisé pour réagencer les caractères sur le clavier en fonction de leur probabilité d'être le prochain caractère à saisir. Ainsi, les caractères les plus probables sont positionnés au plus près du coin supérieur gauche pour limiter le nombre de déplacements que devra effectuer le curseur.

Dans le cas des claviers logiciels utilisés avec un dispositif de pointage pour déplacer un pointeur, le réagencement des caractères n'est pas envisageable car, à la différence du défilement automatique, le pointeur n'est pas repositionné après chaque caractère. Un ré-agencement des caractères imposerait à l'utilisateur un temps de recherche visuel trop important entre la saisie de chaque caractère. L'utilisateur ne pourrait alors plus anticiper les déplacements à effectuer pour aller d'un caractère à un autre. Ce problème n'est pas présent avec un clavier à défilement automatique car l'utilisateur est dépendant du défilement automatique du curseur et ne peut donc pas anticiper la saisie des caractères. Ce temps de défilement peut également lui permettre de regarder la nouvelle disposition sans que cela n'ait d'incidence sur le temps de saisie. Néanmoins, il a été montré avec le système SlideKey [18] que même dans le cadre d'un défilement automatique du curseur, si le système propose un mécanisme de pré-visualisation des résultats à venir, l'utilisateur pourrait valider plus rapidement le caractère suivant lorsque celui-ci se trouve être le plus probable. Le ré-agencement des caractères a montré son efficacité dans un cas spécifique de clavier logiciel manipulé avec un pointeur. Il s'agit des claviers dit « ambigus », tel que le clavier type T9, où plusieurs caractères sont présents sur une même touche. Les différents caractères d'une même touche sont alors accessibles, selon leur position sur la touche, par un, deux ou trois clic(s) sur la touche concernée. Le ré-agencement des caractères est alors bénéfique car d'une part, l'utilisateur n'a que les 3 ou 4 caractères de la touche concernée à regarder ; et d'autre part, le caractère le plus probable sera alors placé en première position sur la touche et sera donc accessible en un seul clic. Cette technique réduit donc le nombre de clic à effectuer et permet ainsi à l'utilisateur de gagner du temps dans sa saisie [21].

Les résultats des algorithmes de prédiction de caractères ont toutefois fait l'objet d'étude sur les claviers logiciels manipulés avec un dispositif de pointage. Magnien et al. [11] ont notamment utilisé la prédiction de caractères pour mettre en évidence les caractères qui ont le plus de chance d'être saisis. Ce système est à destination des utilisateurs qui découvrent une disposition nouvelle des caractères. L'objectif du système est alors d'attirer l'attention de l'utilisateur vers les caractères qui ont le plus de chances d'être saisis.

Mais l'intérêt principal de la prédiction de caractères est de faciliter la phase de sélection du caractère souhaité. D'après la loi de Fitts [6], le temps nécessaire pour pointer une cible est dépendant de la distance pour atteindre cette cible et de la taille de cette dernière. Cette loi est notamment utilisée pour prédire les performances théoriques que pourrait avoir un utilisateur avec un clavier donné [20], et a été utilisée à de nombreuses reprises pour proposer des claviers avec des dispositions de

caractères optimisées pour une saisie au doigt ou avec un pointeur (voir [15] pour une proposition de disposition de caractères optimisée pour le français). D'après la loi de Fitts, il est donc possible d'améliorer la tâche de pointage de deux manières : soit en réduisant la distance à parcourir, soit en augmentant la taille de la cible. Le but est donc d'utiliser les résultats de la prédiction de caractères pour soit diminuer la distance à parcourir pour atteindre le caractère souhaité, soit augmenter la taille des touches contenant les caractères les plus probables. Le système le plus connu qui utilise la prédiction de caractères est le clavier Dasher [24] : tous les caractères sont affichés verticalement sur la droite de l'écran, et défilent de droite à gauche. Pour sélectionner un caractère, l'utilisateur déplace un pointeur sur l'axe vertical au centre de l'écran. Au fur et à mesure de la sélection des caractères, les caractères apparaissent à nouveau sur la droite du clavier. Les résultats de l'algorithme de prédiction de caractères permet de modifier la taille des touches en fonction de la probabilité de saisir le caractère associé à celle-ci. Les caractères les plus probables sont donc sur des touches plus grandes et donc plus facilement sélectionnables. Cependant le défilement automatique de droite à gauche des caractères tout au long de la saisie rend ce système fatigant pour l'utilisateur car il nécessite une attention importante tout au long de la saisie avec ce dispositif. Les claviers BigKey [1] et FloodKey modifient également la taille des touches en fonction de la probabilité d'être saisis des caractères qui y sont associés. Cependant, les touches agrandies par le système BigKey recouvrent en partie les touches voisines. Ceci peut alors engendrer une gêne lors de la sélection de ces touches. De même, les modifications de forme et de taille des touches réalisées par le clavier FloodKey sont apparues trop perturbantes pour les utilisateurs [3]. Le système SpreadKey [14] a l'avantage de conserver la forme initiale du clavier tout au long de la saisie. En revanche, un caractère peut se « propager » sur les touches qui l'entourent lorsque sa probabilité d'être saisi est bien plus importante que celle des caractères des touches voisines. Dans le cas où un caractère est propagé sur les touches voisines, celui-ci peut être saisi sur sa touche d'origine, mais également sur les touches sur lesquelles il s'est propagé. Cependant, chaque caractère reste présent et accessible sur sa touche d'origine quelle que soit sa probabilité d'être saisi. Dans ce cas, le caractère initial est affiché en second plan dans le coin supérieur gauche de la touche et est sélectionnable via un *tap* long ou un geste du pointeur en direction du coin supérieur gauche de la touche. Par conséquent, même si la disposition des caractères changeait, l'utilisateur pourrait quand même effectuer le mouvement qu'il avait anticipé, et ajuster la fin du déplacement du pointeur si le caractère souhaité était propagé sur les touches voisines.

L'autre solution consiste à réduire la distance à parcourir pour atteindre le caractère suivant. Le système KeyGlass [16] propose 4 touches supplémentaires après chaque caractère saisi par l'utilisateur (cf. figure 2). Les 4 caractères les plus probables sont associés à ces nouvelles touches. Les évaluations de ce système ont mis en évidence que les utilisateurs utilisaient bien ces touches supplémentaires (dans

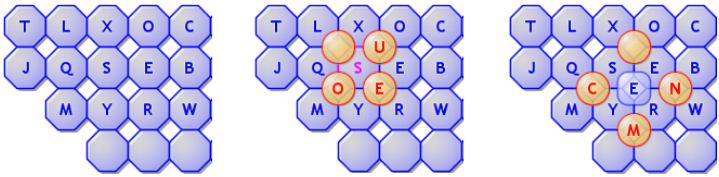


FIGURE 2. Clavier pour la version 2 des KeyGlasses : (gauche) clavier simple ; (centre) KeyGlasses qui apparaissent après la saisie du 's' ; (droite) KeyGlasses qui apparaissent après la saisie du 'e' qui était placé sur une KeyGlass. Extrait de [17].

96 % des cas où le caractère recherché est proposé sur une des touches supplémentaires), et leur utilisation a bien permis de réduire les distances de 53,8 %. En revanche, cette diminution des distances parcourues n'a pas suffi pour augmenter la vitesse de saisie de texte (augmentation du temps pour saisir un mot de 20,1 %). Le fait de proposer aux utilisateurs des caractères supplémentaires les oblige à prendre connaissance de ces nouveaux caractères, et par conséquent les coupe dans leur dynamique, leur anticipation impactant alors négativement leur performance. Cependant, le système a été apprécié des utilisateurs car il permet de minimiser les distances à parcourir et par conséquent réduit la fatigue liée à l'utilisation du clavier logiciel.

Enfin, le clavier sémantique se base sur le principe du pointage sémantique pour améliorer la vitesse de saisie. Le pointage sémantique [5] consiste à moduler le rapport entre le déplacement à l'écran et celui dans l'espace physique en fonction de ce qui se trouve sous le pointeur. S'il n'y a pas d'éléments pertinents sous le pointeur, le déplacement du pointeur est accéléré, et inversement, si le pointeur se trouve sur un objet d'intérêt, sa vitesse de déplacement sera réduite. L'objectif est de faciliter le déplacement d'une cible à une autre et d'améliorer la précision de pointage d'un objet. Ce principe a été repris pour le clavier sémantique en donnant un poids à chaque touche en fonction du caractère qui y est associé et de sa probabilité d'être saisi. La vitesse de déplacement du pointeur sur le clavier est alors liée au poids de chaque touche : plus la touche a un faible poids, plus le pointeur passe rapidement dessus, et plus la touche a un poids élevé, plus le pointeur se déplace lentement sur la touche. L'objectif est ainsi de passer rapidement sur les touches qui ont peu de chance d'être sélectionnées et de faciliter le pointage sur celles dont le caractère a une forte probabilité d'être saisi. Le pointage sémantique a l'avantage de ne pas modifier l'apparence du clavier. Ainsi, par rapport aux autres solutions proposées précédemment, l'utilisateur n'est pas perturbé par les changements dynamiques effectués sur le clavier.

Les systèmes de prédiction linguistique

Que l'on parle de prédiction de caractères ou de prédiction de mots, le rôle de la prédiction linguistique est le même : estimer la probabilité d'occurrence d'un item langagier en fonction de ceux qui sont déjà saisis par l'utilisateur. Cette prédiction fait appel à un modèle de langage qui est estimé, par apprentissage automatique, sur de très grands corpus.

Modèles de prédiction

Dans le domaine de l'aide technique au handicap comme pour la prédiction sur smartphone, on utilise généralement des modèles stochastiques de type $N-$gram compilés sous la forme de transducteurs à états finis [4]. Une fois entraîné, le modèle fournit ainsi la probabilité de l'item linguistique i_k connaissant les $(n-1)$ items de même nature qui le précèdent dans le texte saisi.

Dans le cas de la prédiction de lettres, l'utilisation d'un pentagramme ($N = 5$) est déjà suffisante. Un étude sur le français a par exemple montré qu'ainsi, la lettre recherchée sur un clavier dynamique à 65 caractères permet de présenter la lettre attendue en position moyenne de 2,9 (i.e. ce caractère est en moyenne toujours dans les trois premiers), la maximum théorique de prédiction étant de 2,7 [19]. Les modèles de prédiction de lettres sont très peu gourmands en terme d'occupation mémoire et de temps de prédiction, du fait du nombre limité de caractères. A l'opposé, la prédiction de mots fonctionne sur des dictionnaires de plusieurs dizaines de milliers de mots, ce qui pose le problème de la dimensionnalité des modèles. Le recours à des techniques de compression mais aussi de repliement (*backoff*) pour lutter contre l'éparpillement des données d'apprentissage permet néanmoins d'utiliser là encore des 5-grams de manière efficace. Notons toutefois que les gains de performances en prédiction entre un 3-gram et un 5-gram restent modérés. L'utilisation de techniques d'apprentissage profond pour la prédiction de mots sur dispositifs mobiles fait également face au problème de l'obésité des modèles construits. Une première solution peut être de contourner cette difficulté en déportant les calculs sur un serveur distant, voire de faire appel à des modèles distribués sur le *cloud* [8]. Cette approche est peu appropriée à l'aide au handicap, où la prédiction doit pouvoir opérer en toutes circonstances, en particulier hors ligne. La compression de réseaux récurrents de type LSTM peut toutefois permettre le développement d'une prédiction neuronale efficace implantée sur dispositif mobile [25]. L'absence de jeux d'évaluation partagés rend cependant difficile l'évaluation de l'apport de l'apprentissage profond sur la prédiction de texte : les niveaux de performance rapportées dans le contexte de l'aide à la saisie de texte sont élevés mais restent dans la gamme de résultats observés en prédiction statistique (N-gram). Les temps de calcul rapportés restent par ailleurs de l'ordre du double ou du triple de ceux observés avec des N-grams, sans être a priori pénalisants pour l'utilisateur (de l'ordre de 5 ms par mot).

Évaluation et optimisation de la prédiction de mots

Deux métriques principales sont utilisées pour évaluer l'aide à la saisie de texte. La première est la vitesse de saisie (par mot ou par caractère), qui fournit une évaluation directe de l'aide apportée par le couple modèle prédiction-clavier logiciel. Lorsque l'on cherche à évaluer l'impact de la prédiction seule, la métrique standard est le taux d'économie de saisie (*Keystroke Saving Rate* ou *KSR*). Celui-ci mesure le pourcentage moyen de saisies de caractères (ou d'appuis sur le clavier logiciel) évitées lorsque l'on utilise les propositions correctes de la prédiction pour compléter la saisie. Cette métrique n'est pas totalement indépendante de la disposition du clavier. En particulier, elle dépend du nombre d'hypothèses lexicales présentées à l'utilisateur. Le plus souvent, la mesure retenue est le KSR_5 qui correspond à une liste de prédiction de 5 mots. De nombreuses études ont en effet montré que pour $N = 5$, les courbes de *KSR* commencent à adopter un comportement asymptotique. L'aide à la saisie de texte pour personnes handicapées pouvant concerner des personnes aux capacités motrices et langagières extrêmement variés, il n'existe malheureusement pas de jeu de données de référence représentatif qui serait partagé par toute la communauté. Chaque système étant évalué sur un jeu de test spécifique, il est le plus souvent difficile de comparer les mérites de chaque modèle de prédiction. Toutefois, on note que les systèmes actuels présentent des KSR_5 autour de 0,6 pour des langues telles que l'anglais ou le français [25, 22]. Les langues agglutinatives ou à forte morphologie flexionnelle (allemand par exemple) sont plus délicates à prédire et donnent parfois lieu à des traitements additionnels spécifiques. Sur ces langues, un KSR_5 approchant les 0.5 constitue déjà un bon niveau de performance [22]. Dans tous les cas, on observe ainsi que la prédiction de mots peut éviter plus de la moitié des saisies. Toutes ces études montrent par ailleurs l'importance de développer des modèles de prédiction propre à l'utilisateur en sus de l'apprentissage standard effectué sur de grands jeux de connaissance génériques. Cette adaptation dynamique se fait de manière active au fil des saisies utilisateurs. L'intégration d'une dimension sémantique explicite, sous forme de plongements de mots ou d'analyse sémantique latente, permet également une focalisation thématique de la prédiction. Si elle peut être appréciée qualitativement par l'utilisateur, son impact sur le KSR_5 est toutefois plus limité. Enfin, il est essentiel de noter que le score *KSR* constitue avant tout une estimation théorique de l'aide que peut apporter la prédiction, puisque son calcul suppose que l'utilisateur choisit systématiquement une prédiction lorsqu'elle est correcte. Les études en situation réelle d'usage montrent que c'est rarement le cas, l'utilisateur manquant le plus souvent une bonne prédiction affichée à l'écran. Un des verrous scientifiques centraux de l'aide à la saisie de texte n'est ainsi pas l'amélioration de la prédiction, qui atteint déjà des niveaux de performances très acceptables, mais celle du couplage ergonomique entre prédiction et clavier logiciel. Un autre

enjeu oublié est certainement celui du couplage entre prédiction et correction orthographique.

Prédiction et correction intégrée

La correction orthographique est une question importante pour l'aide à la saisie de texte, à la fois parce que les utilisateurs sont soucieux de la qualité orthographique des textes qu'ils produisent, mais aussi parce que de nombreuses pathologies s'accompagnent de troubles langagiers associés. Les études menées en situation réelle d'usage montrent généralement une amélioration sensible de la qualité orthographique des saisies avec l'emploi d'une prédiction de mots. Souvent l'utilisateur va avoir recours à la prédiction non pas tant pour accélérer la saisie de texte, que pour sélectionner un mot dont il espère qu'il soit ainsi correctement orthographié. Une étude sur 12 mois (non encore publiée) réalisée par le laboratoire LIFAT avec l'hôpital de Garches et le centre de rééducation de Kerpape dans le cadre du projet Predict4All (fondation Bennetot) va prochainement confirmer cette observation. Toutefois, la prédiction peut se retrouver complètement perdue lorsque le texte déjà saisi comporte de trop nombreuses erreurs. Le recours à une correction orthographique intégrée à la prédiction semble alors une nécessité. Malheureusement, on ne dispose pas à l'heure actuelle de correcteurs suffisamment fiables pour répondre à

ce besoin. Une étude [2] menée sur des productions d'enfants paralysés cérébraux ne présentant pas de troubles langagiers a montré que les correcteurs grand public peinaient à corriger la moitié des erreurs présentes. Sur des productions de personnes dyslexiques, ce sont seulement 20 % qui ont pu être corrigées. Dans les autres cas, l'erreur était soit mal corrigée, soit non détectée. Ce problème des non détections concerne majoritairement les erreurs non lexicales, où l'erreur orthographique se traduit par un autre mot de la langue, comme dans l'exemple : *je *mangue une pomme*. Que la personne souffre ou non de trouble dyslexique, ce sont ainsi près de 40 % des erreurs qui ne sont pas identifiées. Les modèles actuels de correction, qu'ils procèdent à une correction hors contexte basée sur une distance d'édition, ou à une correction contextuelle à base d'ensembles de confusion [7], ne sont intrinsèquement pas à même de détecter ce type d'erreurs très fréquentes, même chez le grand public. A notre connaissance, l'application de techniques neuronales n'a jamais été abordée dans le domaine de l'aide à la saisie pour personnes handicapées. Il s'agit donc d'une problématique largement ignorée (voir tout de même [10] dans un contexte proche), qui limite pourtant parfois cruellement l'assistance apportée par la prédiction.

Références

[1] K. Al Faraj, M. Mojahid, and N. Vigouroux. Bigkey : A virtual keyboard for mobile devices. In *International Conference on Human-Computer Interaction*, pages 3–10. Springer, 2009.

[2] J.-Y. Antoine, M. Crochetet, C. Arbizu, E. Lopez, S. Pouplin, A. Besnier, and M. Thebaud. Ma copie adore le vélo : analyse des besoins réels en correction orthographique sur un corpus de dictées d'enfants. In *TALN 2019*, Toulouse, France, 2019.

[3] G. Aulagner, R. François, B. Martin, D. Michel, and M. Raynal. Floodkey : increasing software keyboard keys by reducing needless ones without occultation. In *les actes de la* 10^e *WSEAS international conference on Applied computer science*, pages 412–417. World Scientific and Engineering Academy and Society (WSEAS), 2010.

[4] G. Badr and M. Raynal. WordTree : Results of a word prediction system presented thanks to a tree. In *actes de International Conference on Universal Access in Human-Computer Interaction (LNCS 5616)*, pages 463–471, Berlin, 2009. Springer.

[5] R. Blanch, Y. Guiard, and M. Beaudouin-Lafon. Semantic pointing : improving target acquisition with control-display ratio adaptation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 519–526, 2004.

[6] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6) :381, 1954.

[7] A. R. Golding and D. Roth. A winnow-based approach to context-sensitive spelling correction. *Machine learning*, 34(1) :107–130, 1999.

[8] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv :1811.03604*, 2018.

[9] H. H. Koester and S. Levine. Effect of a word prediction feature on user performance. *Augmentative and alternative communication*, 12(3) :155–168, 1996.

[10] A. Q. Li, L. Sbattella, and R. Tedesco. Polispell : an adaptive spellchecker and predictor for people with dyslexia. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 302–309. Springer, 2013.

[11] L. Magnien, J. L. Bouraoui, and N. Vigouroux. Mobile text input with soft keyboards : optimization by means of visual clues. In *actes de International Conference on Mobile Human-Computer Interaction*, pages 337–341. Springer, 2004.

[12] T. Masui. POBox : An efficient text input method for handheld and ubiquitous computers. In *actes de International Symposium on Handheld and Ubiquitous Computing*, pages 289–300, Berlin, 1999. Springer.

[13] D. Maurel and B. Le Pévédic. The syntactic prediction with token automata : Application to handias system. *Theoretical computer science*, 267(1-2) :121–129, 2001.

[14] B. Merlin and M. Raynal. SpreadKey : Increasing software keyboard key by recycling needless ones. In *actes de 10th European conference for the advancement of assistive technology in europe (AAATE 2009)*, pages 138–143, Amsterdam, 2009. IOS Press.

[15] M. Raynal. Claviers GAG : claviers logiciels optimisés pour la saisie de texte au stylet. In *les actes de la* 18^e *Conference on l'Interaction Homme-Machine*, pages 3–10, 2006.

[16] M. Raynal. Le système keyglass : Système d'ajout dynamique de touches sur clavier logiciel. *Traitement Automatique des Langues, Communication assistée*, 48(2) :97–121, 2007.

[17] M. Raynal. Keyglasses : semi-transparent keys on soft keyboard. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pages 347–349, 2014.

[18] M. Raynal and B. Martin. Slidekey : Impact of in-depth previews for a predictive text entry method. In *International Conference on Computers Helping People with Special Needs*, pages 363–370. Springer, 2020.

[19] I. Schadle, J.-Y. Antoine, B. Le Pévédic, and F. Poirier. Sybilettre, prédiction de lettres pour la communication augmentée, revue d'interaction homme-machine. *RIHM, Revue d'Interaction Homme-Machine*, 3(2) :115–133, 2002.

[20] W. Soukoreff and S. Mackenzie. Theoretical upper and lower bounds on typing speed using a stylus and a soft keyboard. *Behaviour & Information Technology*, 14(6) :370–379, 1995.

[21] N. Vigouroux, F. Vella, P. Truillet, and M. Raynal. Evaluation of aac for text input by two groups of subjects : Able-bodied subjects and disabled motor subjects. In *actes de 8th ERCIM Workshop, User Interface for All, Vienne, Autriche*, pages 28–29, Berlin, 2004. Springer.

[22] T. Wandmacher. *Adaptive word prediction and its application in an assistive communication system*. PhD thesis, Université de Tours, 2008.

[23] T. Wandmacher, J.-Y. Antoine, F. Poirier, and J.-P. Départe. Sibylle, an assistive communication system adapting to the context and its user. *ACM Transactions on Accessible Computing (TACCESS)*, 1(1) :1–30, 2008.

[24] D. J. Ward, A. F. Blackwell, and D. J. MacKay. Dasher : A data entry interface using continuous gestures and language models. In *actes de 13th annual ACM symposium on User interface software and technology – UIST'00*, pages 129–137, New York, 2000. ACM.

[25] S. Yu, N. Kulkarni, H. Lee, and J. Kim. On-device neural language model based word prediction. In *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, pages 128–131, 2018.

La Nuit de l'info

Le bureau de la Nuit de l'info [1]

La Nuit de l'info est une compétition nationale qui, tous les ans, réunit étudiants du supérieur (de bac à bac+8), enseignants et entreprises. Le principe est simple : les participants ont une nuit pour développer une application répondant à un sujet national, tout en relevant des défis complémentaires proposés par des entreprises partenaires.

Pour cette 15e édition, la Nuit de l'info a réuni, dans la nuit du 2 au 3 décembre 2021, environ 3600 participants sur 46 sites répartis dans toute la France mais aussi 6 en Tunisie. Cette édition a été marquée par une motivation certaine des étudiants après une année difficile.

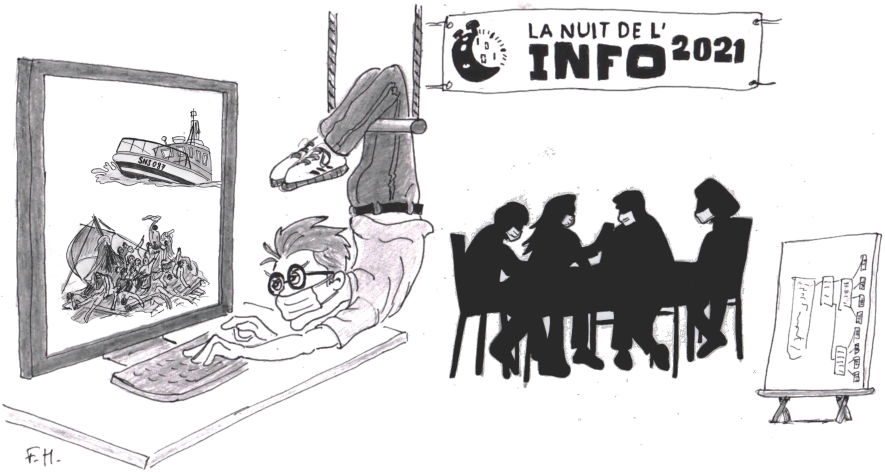
Cette année, les étudiants ont travaillé sur un projet proposé par la Société nationale de sauvetage en mer (SNSM), et plus particulièrement l'antenne de Dunkerque. La SNSM est une association qui œuvre pour le sauvetage en mer, qu'il soit pour les professionnels de la pêche ou les particuliers en détresse, les migrants à la dérive mais également sur les bords de plages l'été avec les maîtres nageurs. Philippe Boutelier y est « sauveteur de mémoire ». Depuis plusieurs années, il recense, classe et met en page les différents sauvetages opérés par la SNSM sur le territoire dunkerquois, du XVIIIe siècle à nos jours, en fouillant les archives de la ville, les journaux anciens ou par le recueil d'informations auprès des internautes ou habitants. Toutes ces informations sont disponibles sur le site de Sauveteurs du dunkerquois [2].

Ces données n'étant pas structurées, il n'est pas possible d'y faire une recherche efficace suivant un nom de famille ou de bateau par exemple. Les participants avaient

1. https://www.nuitdelinfo.com, nuitinfo-bureau@polytech.unice.fr.
2. https://sauveteurdudunkerquois.fr.

DOI:10.48556/SIF.1024.19.27

pour mission de créer une base de données en répertoriant les informations essentielles disponibles sur les pages des « Sauveteurs du dunkerquois ». Ils devaient ensuite développer une interface web permettant la recherche suivant le nom d'un sauveteur ou d'un bateau mais également la possibilité de proposer de nouvelles informations. Malgré un temps imparti très court alloué à ce développement, de nombreuses productions se sont avérées de grande qualité.



La plateforme collaborative qui avait été un élément central lors de la dernière édition de la Nuit de l'info a été une nouvelle fois utilisée pour la communication entre les différentes équipes et les entreprises partenaires.

Un direct a également été diffusé en début de soirée afin de présenter les différents acteurs du projet. Les étudiants ont ainsi fait connaissance avec la SNSM et ont pu poser toutes leurs questions aussi bien sur le rôle de cette association que sur le sujet proposé.

Une des forces de cet événement est la possibilité que les différentes écoles (écoles d'ingénieurs, INP, IUT, BTS, universités...) se mesurent entre elles. Chaque équipe peut choisir cinq défis en plus du défi principal ; ce qui stratégiquement peut sembler une bonne solution (multiplier les chances de gagner) mais cela nécessite une organisation parfaite et une bonne coordination des participants.

Les entreprises partenaires n'ont pas manqué d'imagination en proposant des défis sur les thèmes de l'intelligence artificielle, des NFT, de la minimisation de l'empreinte écologique du site ou encore la mise en valeur des filles dans l'informatique dans un tel projet.

Ce que retiennent les participants de cette expérience ? L'enrichissement des connaissances, gagner en maturité, réaliser le travail en équipe, vivre une expérience

innovante sous contraintes, renforcer les relations enseignants-élèves et le contact avec des professionnels.

Mais cet événement est aussi l'occasion de passer un moment convivial : les équipes rivalisent souvent de créativité sur leurs noms d'équipes, leurs leitmotivs et sur les animations (chaîne vidéo en direct, déguisements, challenges sur les réseaux sociaux, et crêpe party).

Les membres du bureau 2021 de la Nuit de l'info sont : Céline Auzias, Jean-Michel Bruel, Maxime Devanne, Stéphane Isnard, Stéphane Ribas, Julio Santilario Elena, David Roumanet et Rémi Synave et les personnes à l'origine de ce projet sont Pierre-Alain Muller, Sébastien Mosser et Mireille Blay-Fornarino.

ENTRETIEN

Entretien avec Marthe Bonamy, médaille de bronze du CNRS

réalisé par Olivier Baudon [1]

Marthe Bonamy est chargée de recherche au CNRS, affectée au Laboratoire bordelais de recherche en informatique (Labri) de Bordeaux au sein de l'équipe Graphes et optimisation qu'elle dirige depuis 2021. Elle a obtenu la médaille de bronze du CNRS au titre de l'INS2I en 2021. Son principal domaine de recherche est la théorie des graphes dans ses aspects à la fois théoriques et algorithmiques.

O. Baudon, 1024 : « *Peux-tu, tout d'abord, nous rappeler ton parcours ?* »

Marthe Bonamy, MB : Après un master à l'ENS Lyon et une thèse à Montpellier, j'ai été recrutée au CNRS. J'ai décalé mon entrée au CNRS pour faire un post-doctorat au Canada à l'université de Waterloo.

1024 : « *Ton domaine de recherche concerne exclusivement la théorie des graphes ?* »

MB : Je suis vraiment centrée sur les graphes, mais on utilise ou on développe des outils qui ne sont pas vraiment de la théorie des graphes. On est obligé de se tenir au courant. Par exemple : tout ce qui concerne la compression d'informations. Parfois, quand on change la représentation, il faut faire attention à la taille de la nouvelle construction, et on doit réfléchir à la façon de déduire la nouvelle information à partir du moins d'éléments possible. Et cela n'apparaît pas dans la théorie des graphes.

1. Université de Bordeaux.

DOI:10.48556/SIF.1024.19.31

C'est vraiment de la combinatoire pure. Par exemple, récemment, dans un article, on utilise un lemme sur les matrices pour pouvoir compresser une instance 3-SAT en un petit problème de coloration.

1024 : « *Dans ton parcours, qu'est-ce qui t'as amenée vers la théorie des graphes ?* »

MB : C'est un exposé de Louis Esperet. Il y avait une école d'hiver à Lyon, quand j'étais en master 1re année. Toute une équipe de gens de Grenoble était venue. Louis a fait un exposé (qui avec le recul était sans doute assez élémentaire) sur la coloration de graphes. J'étais au premier rang, avec les yeux pleins d'étoiles ! Je suis allé le voir ensuite pour lui demander des pointeurs pour un stage de M1, que je devais faire à l'étranger. Et c'est comme cela que je suis allé faire un stage à Durham au sud de l'Écosse et cela m'a plu.

1024 : « *Donc au départ, ce sont les problèmes de coloration qui t'ont intéressée ?* »

MB : Oui. En fait, ce que j'aime en théorie des graphes, c'est qu'il y a des questions très simples. Même si elles sont en fait difficiles à résoudre. La plupart des problèmes peuvent être abordés en même pas un quart d'heure... Cela suffit pour comprendre la notion.

1024 : « *Oui, le plus classique, c'est le problème des 4 couleurs. Expliquer les 4-couleurs à quelqu'un, c'est facile. Par contre, lui montrer la solution...*

Au CNRS, tu es rattachée à une section qui relève plutôt de l'informatique. C'est volontaire, c'est là où tu pensais avoir le plus de chance d'avoir de la place ? »

MB : C'est là où tous les théoriciens des graphes vont quasiment. La question ne s'est pas vraiment posée pour moi.

1024 : « *Est-ce que tu peux me citer tes principaux résultats (ou concepts) ; ceux dont tu penses qu'ils t'ont valu la médaille de bronze ?* »

MB : D'après ce que j'ai compris, ce qui a joué en ma faveur pour la médaille, ce ne sont pas tant les résultats que le fait que je collabore beaucoup et avec beaucoup de gens différents. C'est cela qu'ils ont apprécié, et que j'aime beaucoup personnellement. Cela m'a fait plaisir qu'ils le remarquent. Dans la façon dont je travaille, je n'ai pas un gros résultat, ou une grosse cathédrale de théorie sur un domaine particulier. Mais, j'aime bien me promener, même si c'est plus difficile en cette période... J'aime bien travailler avec différents groupes, travailler sur leurs problèmes à eux, résoudre leurs trucs (ou pas d'ailleurs)... Je préfère faire plusieurs projets, même si certains demandent beaucoup d'investissement, me diversifier. Notamment me diversifier sur les co-auteurs. J'aime bien voir comment les gens travaillent.

1024 : « *Tu arrives à travailler sur plusieurs sujets en même temps ?* »

MB : Non. Le travail idéal, c'est : une semaine — un projet. J'aime bien avoir quelque chose de bien défini. C'est justement ce qui est difficile avec le confinement. On ne peut pas dire aux gens : « Pas cette semaine, je suis en Allemagne. ». Ça me manque.

1024 : « *Et pour en revenir à l'informatique, quel est ton apport, selon toi, dans ce domaine ?* »

MB : La frontière est toujours floue entre mathématique et informatique. Selon les pays, selon les co-auteurs, ce que je fais est considéré soit comme de l'informatique, soit comme des mathématiques.

1024 : « *Oui, c'est clair que, par exemple en Europe de l'Est, les graphes sont encore considérés comme des mathématiques ; même si cela commence à changer. En Pologne par exemple.* »

MB : Cela dépend des endroits. A Cracovie, par exemple, c'est considéré comme des mathématiques et à Varsovie comme de l'informatique.

1024 : « *Oui, mais par exemple à Cracovie, à l'université de Jagellone, il y a de plus en plus de graphes et c'est dans le département informatique. Et cela leur permet de récupérer de bons étudiants venant de l'informatique.* »

MB : Je pense que ce que cela apporte, c'est surtout des interactions, faire des ponts avec différents outils. C'est cela qui est intéressant. C'est sûr que cela reste perméable entre les deux aspects.

1024 : « *Pour toi, quelles sont les qualités d'une chercheuse ou d'un chercheur, indépendamment de sa discipline ?* »

MB : Je dirais : la remise en question. Je pense que ce n'est pas en restant assis sur ses convictions que l'on avance beaucoup. C'est en s'autorisant à changer d'avis, en s'autorisant à écouter les autres qu'on avance... Mais je pense aussi qu'être têtu, c'est parfois important.

1024 : « *Est-ce que dans l'informatique, il y a des choses, liées éventuellement aux graphes, qui t'ont particulièrement intéressée ?* »

MB : Déjà, dans ce que je fais, il y a pas mal d'algorithmique. Et depuis quelques années, j'essaie de m'intéresser en particulier à l'algorithmique distribuée, mais pour l'instant dans les graphes. Il y a aussi les aspects liés à la compression d'informations : comment décrire quelque chose de façon efficace.

1024 : « *C'est une problématique plus liée à la combinatoire qu'aux graphes directement ?* »

MB : Oui. Mais c'est aussi parfois très appliqué. Typiquement quand tu essaies de représenter des sommets dans des graphes planaires de façon efficace : tu sais

que c'est un graphe à n sommets, mais tu ne sais pas lequel, et tu veux représenter n'importe quel graphe planaire à n sommets en étant sûr de pouvoir le décoder rapidement, que la taille de ta représentation ne soit pas trop grande, etc. Et en fait, derrière ces questions, il y a des questions qui sont à la limite des bases de données par exemple : réussir à stocker une information avec un minimum de place tout en étant capable de la décoder rapidement... Il y a de la théorie des graphes, mais cela ne représente même pas la moitié des intuitions qui sont derrière ces questions.

1024 : « *Qu'est-ce que tu penses de la sous-représentation des femmes en informatique en général, et également dans le milieu de la recherche scientifique (au-delà de la seule discipline informatique) ? Est-ce que tu te sens concernée par cette question ?* »

MB : Forcément, de par le fait que je me prends parfois des commentaires sexistes. Mais au-delà de cela, toutes les jeunes femmes en théorie des graphes que je connais hésitent ou ont hésité à arrêter la recherche, à cause de cela entre autres. Le nombre de doctorantes qui arrêtent la recherche après la thèse est bien plus grand que pour leurs homologues masculins. Et au-delà de savoir combien arrêtent, beaucoup d'entre elles discutent d'arrêter, car cela à l'air d'être un trop gros sacrifice dans leur vie personnelle. Elles ne s'y sentent pas bien, simplement. Je pense que le problème de base, c'est un problème de représentation. Quand quelqu'un pense à une personne forte en mathématique ou en informatique, il ne pense pas à une jeune femme. Et c'est dommage, en particulier pour toutes les jeunes filles qui pourraient contribuer au domaine et que l'on décourage comme cela.

1024 : « *Et tu as suivi un peu les actions de l'association Femmes et informatique ?* »

MB : Un peu.

1024 : « *Et tu en penses quoi ?* »

MB : Je pense que c'est un sujet difficile. C'est dur de quantifier les résultats. Et il faut s'investir, mais il ne faut pas envoyer le mauvais message. C'est aussi cela qui est difficile.

1024 : « *Un message que je fais souvent passer au lycéennes, c'est qu'en informatique, elles auront souvent le choix de l'entreprise où aller travailler car beaucoup sont en demande de plus de mixité, et également que c'est un domaine où elles seront aussi bien payées que les hommes. Mais je suis d'accord avec toi, dans les milieux avec beaucoup d'hommes, il peut y avoir des problèmes sexistes.* »

MB : Je pense aussi qu'en tant que femme, tu n'as pas envie de t'entendre dire que tu vas avoir un truc parce que tu es une femme. Tu as envie de t'entendre dire qu'il y a égalité des chances, que le traitement est neutre, que si tu suis les règles, tout

va bien se passer. Tu n'as pas envie de t'entendre dire « il y a un bonus à la petite jupe ».

1024 : « *Je comprends. Mais c'est vrai que personnellement, quand je devais recruter des étudiants pour une formation, si un garçon était clairement meilleur qu'une fille, je prenais le garçon, mais quand j'arrivais sur les derniers cas, avec des profils équivalents, j'avais tendance à choisir les filles pour équilibrer un peu.* »

MB : Oui, mais ce sont des choses à faire, pas à dire !

1024 : « *Un autre point qui me pose problème, ce sont les règles de parité, par exemple dans les comités de sélection. Cela part d'une bonne idée, mais je vois que les collègues femmes professeures sont complètement sous l'eau au moment des comités de sélection car elles sont demandées partout.* »

MB : Elles peuvent dire non.

1024 : « *C'est difficile de dire non à tout le monde. Tu peux dire non quand tu as déjà dit oui à un autre...* »

MB : Effectivement, quelque chose que je constate chez beaucoup de mes collègues féminines, c'est que nous avons du mal à dire non. Moi, cette année, j'ai fait trois recrutements MCF, c'est trop. Ça m'a pas mal plombée. Je pense qu'il faut qu'on arrive à dire non plus souvent. Si les gens ont du mal à faire leur jury, c'est qu'il faut plus de femmes ! Ce n'est pas à nous de payer le prix de cette règle. La raison des quotas des jurys, ce n'est pas pour les membres, c'est parce que c'est bien pour les candidats. Ce n'est pas pour qu'il y ait plus de femmes, mais parce que les candidates seront plus à l'aise et que les membres masculins du jury vont avoir tendance à faire un peu plus attention à ce qu'ils disent.

1024 : « *Est-ce que tu connais la SIF ? Qu'est-ce que tu en penses ?* »

MB : Oui, je connais. C'est toujours bien d'organiser une discipline en France. Et dans tous les pays d'ailleurs. Dans tous les pays que je connais, il y a une association pour l'informatique, les mathématiques. Après, chacun voit son rôle différemment, mais c'est toujours bien d'avoir quelque chose qui représente la discipline dans le pays.

1024 : « *Comme tu as l'occasion de visiter beaucoup de pays, est-ce que tu as un avis sur l'enseignement de l'informatique, sur la recherche en France ou la formation doctorale... ?* »

MB : Cela varie énormément selon les pays, selon les domaines aussi bien sûr. Si on parle juste de l'informatique, ou même de la théorie des graphes, par exemple, ce qui varie beaucoup, c'est l'importance qui est donnée aux doctorants. Il y a des pays où le doctorant est un sous-fifre qui fera ce qu'on lui demande et s'il est gentil, il aura son doctorat à la fin. Je pense que la France est un des pays où le doctorant est le mieux considéré. Et encore, cela dépend des équipes, il y a encore pas mal de facteurs. Mais on a plus tendance à traiter les doctorants comme des collègues. Je trouve cela assez dommage quand je vois des doctorants qui n'ont pas beaucoup de liberté, pour lesquels certaines tâches, comme la correction de copies, mangent leur temps de recherche. Certes, cela les prépare à leur métier d'enseignant-chercheur, mais il ne faut pas non plus qu'ils soient les employés mal payés... Je pense que c'est surtout une question de respect envers les jeunes qui arrivent : les prendre au sérieux, les voir comme les futurs chercheurs et pas comme de la main d'œuvre facilement disponible.

1024 : « *Tu as déjà encadré un certain nombre de doctorants. Je suppose que tu aimes ça. Ce que tu aimes bien, c'est l'interaction ?* »

MB : C'est l'interaction. Mais c'est aussi voir l'évolution au fil des années. C'est extrêmement gratifiant quand on voit son ancien doctorant — dont on se souvient comme étant mal à l'aise, timide... — tout à coup capable de gérer un projet sans aucun problème, amorcer des collaborations, avoir des super idées pour un projet. J'aime beaucoup l'interaction, mais j'aime surtout voir le chemin qui a été fait.

1024 : « *Ton avis concernant les différents statuts de chercheurs en France : tu souhaitais en priorité être recrutée au CNRS, plutôt que maître de conférences ou à Inria par exemple ?* »

MB : Oui, pour la flexibilité. Au CNRS, c'est très appréciable de pouvoir faire des cours si on veut, mais pas d'être contraint à en faire. Il y a des semestres où l'on souhaite se concentrer sur autre chose. Et aussi pour la mobilité. Si pour des raisons personnelles, je dois aller à un autre endroit, c'est possible. C'est quelque chose qui me plaît. Si je n'avais pas eu un poste au CNRS, j'aurais clairement candidaté à d'autres types de postes. Mais c'était clairement mon choix numéro un, et je ne m'en plains toujours pas.

1024 : « *Personnellement, quand j'ai eu ma thèse, je n'ai candidaté que sur des postes de MCF, car j'ai fait ma thèse dans un laboratoire avec une majorité de chercheurs CNRS, et je voyais souvent certains d'entre eux qui avaient fini de traiter un sujet ou qui bloquaient et qui en manque d'idée avaient tendance à déprimer un peu. Alors que l'enseignement permet de garder un rythme et de se sentir utile quelles que soit les avancées en recherche. Tu n'as jamais eu ce problème ?* »

MB : Non, je n'ai jamais rencontré ce problème. Évidemment il y a des moments où je me sens plus à l'aise dans mon métier que dans d'autres, il y a des moments où je me demande pourquoi le CNRS m'a recrutée (rires...) mais aussi d'autres moments où ça va, où j'arrive à faire des choses. Mais même dans les moments de creux où je n'arrive à rien, où j'ai l'impression de ne pas faire la bonne recherche, ne pas réussir à prouver ce que je veux, je ne suis quand même pas bloquée. J'ai toujours quelques hypothèses à tester, 450 mails à traiter, il y a toujours moyen de s'occuper. C'est plutôt l'inverse. C'est plus quand je passe trop de temps à faire de la recherche que je culpabilise de ne pas faire les autres choses que j'étais censée faire, typiquement répondre à mes mails, ou gérer les aspects administratifs, les invités, les projets...

1024 : « *Tu enseignes en M2 ?* »

MB : J'enseigne en M2, j'ai un cours à Bordeaux et un cours à Lyon.

1024 : « *Tu trouves cela intéressant, important ?* »

MB : Oui, c'est intéressant, c'est stimulant. Et puis du coup, il y des choses que je comprends mieux maintenant que j'ai dû les expliquer. Il y a des parties de la théorie des graphes que je connaissais, mais pas profondément avant d'avoir à les expliquer à des étudiants. Après, cette année, c'est en distanciel. C'est donc différent. Faire son cours avec des transparents en ligne, ce n'est pas très gratifiant. Mais je suis assez contente en général.

1024 : « *Les cours que tu donnes sont vraiment dans ta spécialité.* »

MB : Oui c'est sûr. Mais on peut sortir de cours en ayant l'impression de ne pas avoir expliqué clairement les choses. Typiquement, si en sortant du cours, un étudiant me

demande la définition qui était sur le transparent 1, je me dis que j'ai un peu raté ma séance. Bon, ce n'est pas arrivé, heureusement. En ligne, c'est difficile de voir si les étudiants ont suivi ou pas. En présentiel, on peut plus facilement s'adapter, donner des exemples, etc. En plus, faire des dessins au *touchpad* sur l'écran BigBlueButton, ce n'est pas l'expérience la plus sympa, que ce soit pour les étudiants ou pour moi. Mais ils ont l'air de s'en être bien sortis, je suis restée en contact avec des anciens étudiants, que ce soit à Bordeaux ou à Lyon. Je me dis que je ne les ai pas suffisamment traumatisés pour qu'ils n'aient pas envie de me parler.

1024 : « *C'est quoi pour toi un bon laboratoire ? Quelles qualités tu attends d'un bon laboratoire ?* »

MB : Je pense que c'est important qu'il y ait une bonne ambiance dans l'équipe, qu'il y ait un bon accueil des nouveaux arrivants. Cela fait une grosse différence à l'entrée si personne ne te parle ou si des gens viennent te voir en te demandant comment cela se passe, en te donnant des suggestions, etc. Et au-delà de l'équipe, c'est important qu'il y ait une bonne ambiance entre les équipes, pas de piques entre les équipes pour savoir qui fait de la vraie science. Et aussi que l'on fasse confiance aux équipes, qu'on leur donne les moyens de faire les choses correctement. J'en suis à mon troisième directeur de laboratoire depuis mon arrivée. Et à chaque fois, j'ai noté un respect pour la théorie des graphes. Donc tout va bien. Je pense que c'est gênant quand la direction se laisse prendre dans des guerres entre les équipes.

Ce que j'aime bien au Labri, c'est que, comme je fais l'accueil tous les ans des L3 de l'ENS Lyon, cela me force à connaître pas mal de gens, de toutes les équipes. J'apprécie. Quand je me promène dans les couloirs, c'est rare que je croise des gens que je ne connais pas, à part des doctorants : ils changent trop vite pour que je mémorise tous leurs noms.

Exposition « Des Elles pour le Numérique »

Olivier Baudon [1]

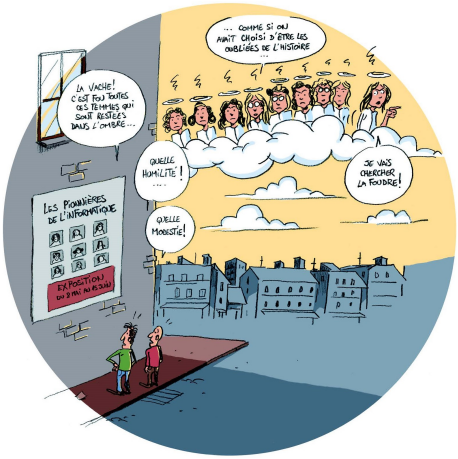
Cette exposition est née suite à un souhait du comité d'organisation de la Robocup à Bordeaux. Sa réalisation a été pilotée par Marion Paoletti, chargée de mission « Parité, égalité, diversité » à l'université de Bordeaux. Elle a été présentée à partir du 8 novembre 2021 à l'IUT de Bordeaux, ainsi qu'au LaBRI (Laboratoire bordelais de recherche en informatique) lors d'un exposé donné par Isabelle Collet sur « Femmes et informatique : pratiques égalitaires, dispositifs inclusifs » le 9 novembre 2021. Les dessins sont disponibles au format A5 et font l'objet d'expositions itinérantes pour sensibiliser aux biais liés au manque de parité dans le numérique. L'ensemble des dessins a été publié dans un bulletin — disponible sur le site de la SIF [2] *(l'un des dessins est reproduit ci-après avec l'autorisation de l'association) — dont nous reproduisons ci-dessous le propos liminaire.*

« Cette exposition est le fruit d'une collaboration entre Bordeaux Métropole, l'université de Bordeaux et l'atelier Croc en Jambe, un collectif d'auteurs et d'autrices de bandes dessinées créé en 2006 et basé à Bordeaux. La volonté de ce travail collaboratif est de montrer les biais sexistes qui persistent dans le milieu du numérique et de l'intelligence artificielle, la quasi-absence des femmes et ses conséquences. Ce livret s'inspire d'un ouvrage d'Aude Bernheim et Flora Vincent paru en

1. Université de Bordeaux.

2. http://www.societe-informatique-de-france.fr/2021/12/des-elles-pour-le-numerique/.

DOI:10.48556/SIF.1024.19.39



2019 : *L'intelligence artificielle, pas sans elles !* Ce travail s'appuie sur un ensemble de résultats pour mettre en lumière les biais, la perpétuation des stéréotypes et les conséquences qui en découlent.

On pense à tort que les ordinateurs sont des solutions neutres pour supprimer les inégalités et prendre des décisions objectives. En effet, ils ont des capacités qui surpassent parfois les nôtres. Pour autant, les programmes sont élaborés par les humains et reproduisent finalement les biais, stéréotypes et représentations collectives, au risque de les ancrer par la technique sous couvert de neutralité.

Aujourd'hui, l'intelligence artificielle prend de plus en plus de place dans nos vies, sans que nous nous en rendions toujours compte. Elle aide à la prise de décision dans de nombreux domaines qui nous concernent tous et toutes, domaines desquels les femmes ne peuvent ni ne doivent être exclues : santé, éducation, justice, travail...

L'exposition est articulée autour de différents thèmes. Ces thèmes n'ont pas été choisis au hasard : ils reflètent les enjeux auxquels fait face le monde du numérique aujourd'hui. Ce domaine est reconnu pour être principalement porté par des hommes et dans une perspective masculine. On est parfois loin de percevoir les mécanismes par lesquels les femmes et le point de vue féminin peuvent être exclus, ainsi que leurs conséquences. Avec humour et sagacité, ce livret apporte un éclairage sur les raisons pour lesquelles les femmes sont plutôt absentes du numérique et l'intérêt de développer leur présence. »

Les mini-ordinateurs « Éducation nationale » de la décennie 1970

Daniel Caous [1] et Jacques Baudé [2]

Cet article présente les matériels ayant fait l'objet de cahiers des charges « Éducation nationale » et qui ont équipé 58 lycées de 1973 à 1976 : il s'agit des deux mini-ordinateurs CII Mitra 15 et Télémécanique T1600 choisis pour l'opération dite « Expérience des 58 lycées [3] ». Aux informations techniques se mêlent les souvenirs vécus des auteurs.

Pour cette opération informatique expérimentale menée à partir de 1973, et parmi moult établissements de l'enseignement secondaire intéressés et candidats, 58 d'entre-eux furent retenus. Le cahier des charges présenté aux constructeurs d'ordinateurs était serré, tant en termes de délais que d'exigences techniques ; l'implémentation du langage français de programmation LSE de SupÉlec était exigée. Au final, parmi 5 fabricants présélectionnés initialement, il ne fut fait appel qu'à deux d'entre-eux : Compagnie internationale de l'informatique (CII) pour le Mitra 15, et Télémécanique (sa division de l'informatique industrielle) pour le T1600 [4].

1. Ex-élève « 58 lycées » de 1976 à 1981 au lycée de Bréquigny à Rennes, chef de projet à la direction du numérique et des systèmes d'information au conseil régional de Bretagne.

2. Ex-secrétaire général et président de l'EPI de 1981 à 1995, ex-enseignant de l'option expérimentale d'informatique de la décennie 1970, président d'honneur de l'EPI, membre d'honneur de la SiF.

3. L'expérience des « 58 lycées », Jacques Baudé, 1024, https://doi.org/10.48556/SIF.1024.4.105.

4. La liste des 58 lycées, avec pour chacun d'entre eux l'année d'équipement et le modèle de mini-ordinateur installé : https://www.epi.asso.fr/revue/histo/h70-58lycees.htm.

DOI:10.48556/SIF.1024.19.41

Un des premiers établissements concernés fut le lycée de La Celle Saint-Cloud, en février 1973, équipé avec le Mitra 15 ayant séjourné à SupÉlec pour la mise au point du LSE. Cet ordinateur prenait la succession d'un 10 010 de la CII préalablement expérimenté dans l'établissement[3].

Ces nouveaux mini-ordinateurs Mitra 15 et T1600 sont livrés pour le premier en configuration « armoire à tiroirs modulables » et pour le second en version « armoire industrielle », tous deux avec pupitre de commandes en façade. Ils disposent d'unités centrales à mots de 16 bits, avec une mémoire vive de 8 à 16 Koctets, à tores de ferrite. Le moniteur en « temps partagé » mis à disposition sur ces machines permet le fonctionnement de la configuration livrée en standard : 8 consoles de visualisation Sintra TTE avec clavier américain QWERTY et un téléimprimeur Teletype ASR-33 avec lecteur-perforateur de ruban (décliné en deux versions carrossées différemment : celle livrée avec le Télémécanique T1600 étant insonorisée)[5].

Le système LSE[6] livré sur chacun des deux mini-ordinateurs, assure une portabilité bien supérieure à ce qu'aurait permis alors toute autre choix. Le Teletype ASR-33 — un standard de robustesse reconnu dans le monde — sert à la fois de poste de contrôle du système, d'imprimante et de moyen de sauvegarde de fichiers grâce au lecteur-perforateur de ruban intégré. La présence de son clavier en fait ponctuellement un terminal supplémentaire d'appoint (fort peu confortable, cependant), à des fins d'écriture de programmes. En 1976, un lecteur de disque souple (*floppy disk*) — interne et intégré en façade sur les Mitra 15, externe et aussi volumineux qu'un lave-vaisselle pour les T1600, logeant des disquettes de taille... 8 pouces, de capacité limitée à 128 Ko à l'époque ! — viendra compléter chacun des 58 ensembles informatiques. Le disque dur interne livré de base, encombrant, presque 1 mètre de côté, autant de hauteur et profondeur, d'une capacité de 256 Ko, s'avère très vite très insuffisant ; cette limitation imposera le doublement de la capacité du disque dur, voire davantage, selon les moyens des établissements. Également, le Teletype ASR-33 avait un fonctionnement mécanique bruyant et plutôt lent : son débit alignait au mieux 15 caractères à la seconde en impression simple, vitesse rabaissée à ... 10 caractères à la seconde lors d'une impression avec lecture ou perforation de ruban ! C'est pourquoi quelques établissements s'équipèrent ultérieurement soit d'un téléimprimeur rapide, soit d'une imprimante matricielle performante. Enfin, selon les spécificités des matières à enseigner, plusieurs de ces 58 configurations furent complétées de périphériques divers : second lecteur de disquettes, consoles graphiques, terminaux clavier-écran supplémentaires. Exemple : le lycée Rive Gauche / Le Mirail à Toulouse, doté initialement d'un ensemble informatique standard à base de

5. Photo d'assemblage des matériels ordinateurs, consoles de visualisation, téléscripteurs de l'opération « 58 lycées ». Extrait du diaporama de Pierre Ratinaud, maître de conférences à l'université de Toulouse 2-Le Mirail : https://slideplayer.fr/slide/489391/1/images/5/Des+enseignants+reçoivent+une+formation+longue+en+informatique+(1+an).jpg.

6. Le système LSE, Jacques Baudé, 1024, https://doi.org/10.48556/SIF.1024.7.41.

CII Mitra 15 avec disque dur de 400 Ko, 8 terminaux texte-clavier écran Sintra TTE et un téléimprimeur Teletype ASR-33 : la configuration fut complétée plus tard par une console graphique Tektronix ainsi qu'une imprimante matricielle rapide Mannesmann Tally. [7].

Parallèlement aux 58 établissements du secondaire équipés, il est à noter que d'autres mini-ordinateurs CII Mitra 15 et Télémécanique T1600, livrés avec divers langages de programmation implémentés (APL, Basic, Fortran, LSE, Lisp, PL/1), furent installés dans divers secteurs : unités de recherche de l'enseignement universitaire, industrie, énergie, centraux de télécommunications, supervision du transport ferroviaire, places militaires de la Défense nationale, ministères. Ces ordinateurs performants restèrent d'ailleurs en service pour certains d'entre-eux jusque... dans les années 2000 ! (voir webographie).

Télémécanique T1600

Historiquement, en 1971, la division d'informatique industrielle de Télémécanique avait conçu en maquette l'ordinateur T800 destiné à succéder à ses T1000 et T2000, ces deux modèles étant installés dans les centraux de supervision des Télécommunications. Face aux exigences du cahier des charges « 58 lycées » du ministère de l'Éducation nationale, et pour remporter le marché, la décision de Télémécanique fut de doubler la capacité et les performances de son prototype T800. Cette décision aboutit au T1600 (cf. ci-contre) : ordinateur final, nouveau, puissant, à mots de 16 bits, qui reprenait l'apparence robuste du T2000, conformément à ce que savait fabriquer le constructeur. Du côté du langage de programmation LSE de SupÉlec, le choix de Télémécanique fut d'en sous-traiter la mise au point à une entreprise externe de services (démarche différente de celle adoptée dans le même temps par CII et SupÉlec, pour la mise au point du langage LSE sur le Mitra 15). Le mini-ordinateur T1600 construit et livré en configuration d'armoire industrielle, abritait l'unité centrale, le disque dur de 256 Koctets, les cartes de mémoire pour un total de... 16 Koctets (extensible), et les interfaces. Il répondait exactement aux spécificités requises par le cahier des

T1600 du lycée Joffre à Montpellier.

7. Galerie de photos des configurations « 58 lycées » de deux établissements de l'enseignement secondaire à Toulouse : le lycée Rive Gauche / Le Mirail équipé avec un Mitra 15, et le lycée Saint-Sernin doté d'un T1600. Collection de Jean-Daniel Dodin et de Mary-Denise Dodin, enseignants au lycée Rive Gauche Le Mirail à Toulouse, lors de l'époque « 58 lycées » : `http://dodin.org/piwigo/index.php?/tags/230-mitra_15_informatique`.

charges de l'Éducation nationale, d'autant que le LSE, exigé, y était implémenté, bien entendu [8].

Grâce à la zone de *swap* disque utilisée par la mémoire centrale, le moniteur en temps partagé livré avec le T1600 permettait d'utiliser en mode conversationnel et simultané plusieurs terminaux clavier-écran ; tout ceci était matérialisé par des cartes et coupleurs d'interfaçage installés en châssis, dans l'armoire de l'ordinateur. Cette électronique interne, via des gros câbles de données adressait le disque dur, le Teletype ASR-33, chacune des consoles de visualisation Sintra TTE, et par la suite le lecteur externe de disquettes 8 pouces.

L'ordinateur Télémécanique T1600 retenu fut installé dès 1973 dans 31 des 58 établissements de l'enseignement secondaire, dont le lycée de Bréquigny à Rennes, en 1974, le CII Mitra 15 équipa quant à lui les 27 autres. Côté langage de programmation, sur l'ordinateur T1600, il est à noter qu'à partir de 1978, LSE implémenté évolua plusieurs fois dans son ergonomie et ses fonctionnalités d'accès aux périphériques. Évolution non pas grâce au constructeur Télémécanique, mais par l'initiative de plusieurs enseignants contributeurs, en mode collaboratif : ceux-ci, probablement spécialisés « ingénierie système », étaient dispersés dans certains des « 58 lycées » dotés d'un T1600, établissements porteurs situés en région parisienne et dans le sud-ouest du pays.

Dans la salle de l'ordinateur, au sein du lycée rennais, cette technologie moderne, toute neuve, impressionnante, faisait planer une atmosphère quasi mystique, parmi les lycéens du club informatique, tous plus passionnés les uns que les autres [9].

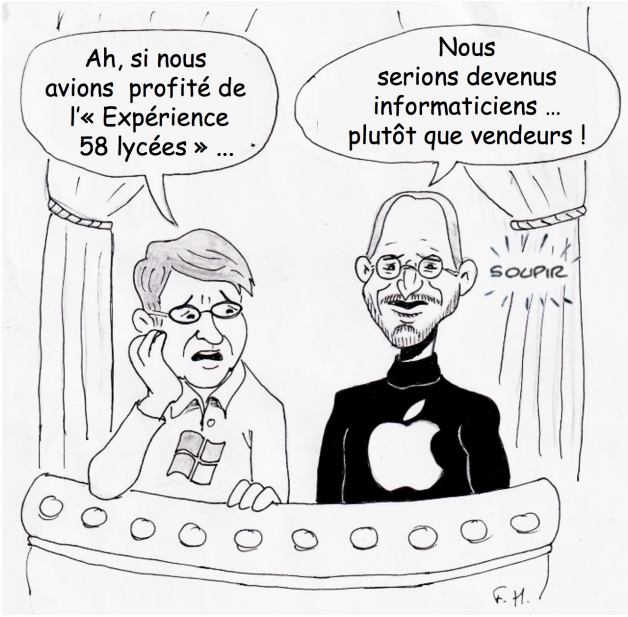
La gestion technique d'une telle salle était réservée aux enseignants « de référence » ; mais au fil du temps et dans divers lycées, elle fut également déléguée à quelques élèves « de confiance » du club informatique (lesquels ne demandaient que ça...), selon les compatibilités, obligations et vacances d'emplois du temps des uns et des autres [10].

Cette délégation de confiance à des élèves pour superviser la salle informatique présentait au moins quatre avantages. D'abord, pour permettre un meilleur accès du lieu à d'autres enseignants ou élèves, ceux-ci pas spécialement au fait des procédures de démarrage et arrêt de l'ordinateur : des utilisateurs soit « apprentis programmeurs », soit professeurs encadrant une classe « qui venait à l'ordinateur » (sic) pour utiliser les logiciels déjà écrits. Ensuite, pour programmer en « libre service »,

8. L'étude « pré 58 lycées » de MM. Daniel Quéniart et Jean-Michel Yolin en 1971 pour le ministère de l'Éducation nationale : `https://www.epi.asso.fr/blocnote/etude_queniart-yolin.pdf`. Galerie de photos de l'ordinateur T1600 qui a équipé le laboratoire Lactamme de l'École polytechnique en 1972 : `http://www.lactamme.polytechnique.fr/images/T1600.21.D/display.html`.

9. Témoignage du parcours de lycéen dans les méandres passionnantes d'un club informatique des années 1975 dans l'un des « 58 lycées », à Rennes : `https://www.epi.asso.fr/revue/histo/h75-lse-caous20.htm`.

10. Entretien avec un développeur de logiciels, ancien élève « 58 lycées », à Albi, année 1975 , `https://cpcrulez.fr/auteur-michel_martin.htm`.



dans ce cadre privilégié d'enseignement-acquisition d'un langage de programmation évolué : LSE, langage de programmation en français. Également, cela cassait les codes de hiérarchie « enseignant-élève » — juste dans ce contexte précis — et plaçait tous les acteurs à un échelon presque identique : celui des passionnés, amateurs d'informatique (loisir complètement incongru, décalé à l'époque), créateurs, concepteurs (pour qui « tout était à construire » !) ; le partage de ces nouvelles valeurs technologiques y était le credo permanent, avec bien entendu les « sachants » et les « apprenants », sous l'égide du Dieu Ordinateur... Enfin, cela permettait aux jeunes néophytes délégataires d'acquérir à leur tour l'expérience pré-professionnelle de ce qui s'avère être aujourd'hui en 2022 l'administration de premier niveau d'une plate-forme informatique.

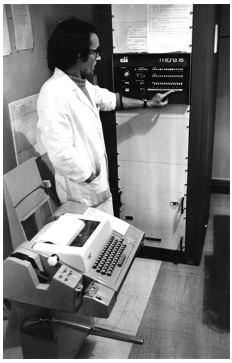
Lors des diverses sessions, la mise en service du T1600 était relativement simple (rôle prestigieux pour les rares élèves autorisés à effectuer cette tâche) : sur l'armoire de distribution électrique, mise sous tension de divers interrupteurs ; puis sur le pupitre frontal de l'ordinateur, par l'action des touches de commande : ST / Sous-tension, AR / Arrêt, INI / Initialisation, MA / Marche ; enfin, dialogue conversationnel sur le téléscripteur (console maître), avec quelques questions techniques posées par le système, et les réponses exactes (Oui/Non) à fournir au clavier. La totalité de la configuration informatique devenait alors magiquement opérationnelle, avec l'ambiance et le décor appropriés : divers voyants allumés ou clignotants sur le pupitre

de l'ordinateur, mise en service des consoles de visualisation, atmosphère affairée et studieuse élèves professeurs, créativité avec échanges de thèmes sur les programmes (à écrire !), lignes de commandes et de programmation affichées sur les écrans, bruit du téléscripteur lors de l'impression des listings. Cette ruche de modernité, berceau technologique des savoirs nouveaux, lieu exclusif de distribution-acquisition de telles connaissances, évoquait sans difficulté les décors de séries de science-fiction diffusées dans le même temps à la télévision...

CII Mitra 15

Pour la mise au point de LSE sur le CII Mitra 15, la démarche s'opère différemment du choix de sous-traitance de Télémécanique pour son T1600. L'implémentation de LSE sur CII Mitra 15 se fait par le biais d'une coopération entre l'équipe des ingénieurs de Jacques Hebenstreit, dont Yves Noyelle et Stéphane Berche, de l'École supérieure d'électricité, et l'équipe enseignante du lycée-collège de La Celle Saint-Cloud. Cette équipe enseignante pluridisciplinaire (mathématique, physique, biologie, histoire-géographie, anglais, musique) avait déjà par ailleurs une sérieuse expérience des exigences et contraintes liées à l'enseignement de l'informatique dans le second cycle qui avait débuté en 1970. Je rappelle l'origine de cette expérience originale peu connue : la Compagnie internationale pour l'informatique (CII) étant relativement proche de La Celle Saint-Cloud, nombre de parents d'élèves y travaillaient. À la suite d'une visite organisée début 1969 pour des élèves de 1ère C et quelques professeurs, des parents présents au Conseil d'établissement du lycée proposèrent le prêt d'un ordinateur : un 10 010 [11] accompagné de 4 Teletypes ASR-33 et un téléscripteur-maître avec lecteur « rapide » de ruban perforé (cf. figures page suivante). Le langage utilisé était Fortran.

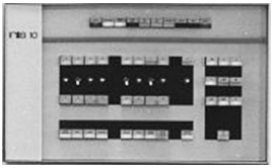
Ce fut cette situation favorable qui permit l'expérience pionnière d'enseignement de l'informatique pilotée par l'Institut national de recherche et de documentation pédagogique (INRDP) et l'Office français des techniques modernes de l'Éducation (OFRATEME).



Remarquons que si on a coutume de dire que l'introduction de l'informatique dans l'enseignement français trouve son origine dans le séminaire de Sèvres (mars 1970) la décision de la CII de prêter un ordinateur au lycée-collège de La Celle Saint-Cloud, et celle de l'INRDP de créer une option expérimentale, datent de 1969 [12].

11. Le 10 010, http://www.feb-patrimoine.com/projet/10010/CII_10010.htm.
12. Le séminaire de Sèvres, Jacques Baudé, 1024, https://doi.org/10.48556/SIF.1024.11.115.

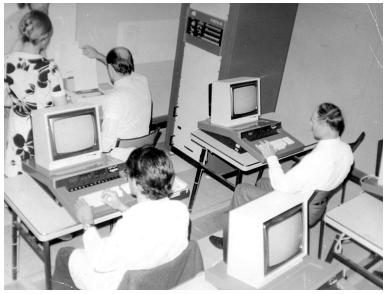
Le lycée-collège fut donc naturellement prioritaire pour la dotation d'un Mitra 15 de la CII qui arriva en février 1973 et l'enseignement (option informatique expérimentale et utilisation de l'informatique dans les différentes disciplines) put continuer en LSE. (Photo : le Mitra 15, ASR-33 et consoles Sintra au lycée-collège de La Celle Saint-Cloud en 1973).



(a) IRIS 10 (10.010 réorienté)

(b) ASR-33

Nous ne revenons pas sur les caractéristiques techniques déjà traitées précédemment : Mini-ordinateur 16 bits, mémoire centrale à tores magnétiques, disque dur de 256 Ko qui s'est vite révélé insuffisant, temps partagé, entrée par un lecteur électromécanique de ruban perforé sur l'ASR-33 (temps de chargement de LSE, en cas de panne, plus d'une heure !). Il faudra attendre 1977 pour être doté d'un floppy disque... avec en tout huit consoles de visualisation [13].



Les débuts furent difficiles. Le LSE n'était pas totalement au point d'où des pannes système. Nous devions téléphoner à Yves Noyelle en lui donnant l'état des voyants. Il arrivait rapidement de SupElec avec son volumineux listing sous le bras et se livrait à quelques opérations salvatrices sur les clés. Mais il y eu aussi des pannes matérielles dues notamment aux variations de tension du courant. Je me souviens de l'installation d'un « mouchard » pendant plusieurs jours pour convaincre l'EDF que le courant chutait brutalement de 220 volts à 180 volts, ce qui n'était pas du goût des têtes de lecture du disque dur !

J'avais appris à prévoir un TP de biologie de secours quand j'amenais travailler une classe de terminale sur un logiciel de simulation, Mendel, Linkover, Glycm ou autres. Mais les aspects positifs l'emportaient. Il faut se souvenir que dans les années 70 aucun élève n'avait accès à un ordinateur en dehors des rares lycées équipés. Nous étions des privilégiés. Tout était à inventer : l'enseignement de l'informatique à ce niveau, la programmation en LSE (après le Fortran), l'encadrement du club informatique et, en parallèle, la conception de programmes pédagogiques dans le

13. Voir également pour le Mitra 15 : https://fr.wikipedia.org/wiki/Mitra_15, https://www.feb-patrimoine.com/projet/mitra/mitra.htm

cadre des groupes disciplinaires de l'INRP [14]. Pour tout cela, nous avions quelques heures de décharge de service qui ne suffisaient évidemment pas mais les pionniers volontaires s'en accommodaient !

Conclusion

L'opération « 58 lycées » fut une des composantes de la réponse française aux conclusions du séminaire de Sèvres [15]. Une autre composante importante a été la formation « lourde » des enseignants (plus de 500, de 1970 à 1976). Ainsi, la formation précédait les matériels et déterminait le choix des établissements à équiper.

Les conditions étaient réunies pour faire de cette opération une réussite. La rencontre française de deux mini-ordinateurs Mitra 15 et T1600 et du système LSE assurait compatibilité et portabilité excellentes [16]. Mais néanmoins réussite partielle car le Comité pédagogique a donné la place prépondérante à l'outil pédagogique dans les disciplines générales.

Heureusement, un certain enseignement de l'informatique, dont la programmation, s'est développé à l'initiative des enseignants dans les clubs ouverts dans les lycées équipés. Le concours de programme AFCET [17] a montré ce que de jeunes élèves étaient capables de réaliser en matière d'inventivité, de programmation et de dossier d'accompagnement. Sur les 103 dossiers retenus, 31 avaient comme support le Mitra 15 et 28 le T1600.

Mais, pour la création d'une option informatique, il faudra attendre la décennie 1980 [18]... après la clôture de l'opération « 58 lycées » !

Parallèlement, beaucoup d'enseignants formés se sont impliqués dans les groupes pédagogiques disciplinaires de l'INRP et ont créé des logiciels pédagogiques encore utilisés sur PC les décennies suivantes[14].

Cette période pionnière a été singulière dans les premiers déploiements de l'informatique pédagogique dans notre pays. Riche d'enseignements et de compétences, elle marquera durablement les esprits et les développements à venir.

---

14. Dix ans d'informatique dans l'enseignement secondaire (1970-1980) INRP, 182 pages, https://www.epi.asso.fr/blocnote/Dix_ans_INRP_1981.pdf.

15. Le séminaire de Sèvres, Jacques Baudé, 1024, https://doi.org/10.48556/SIF.1024.11.115.

16. Un incontournable : « Pour une histoire de l'informatique dans l'enseignement français, premiers jalons » par Émilien Pélisset, ex-président de l'EPI, dans Système éducatif et révolution informatique (Cahiers de la FEN - 1985) : https://edutice.archives-ouvertes.fr/edutice-00284085/file/index.html.

17. Concours de programmes AFCET — 1981, Jacques Baudé, 1024, https://doi.org/10.48556/SIF.1024.18.117.

18. L'option informatique des lycées dans les années 80 et 90, Jacques Baudé, 1024, https://doi.org/10.48556/SIF.1024.2.85, https://edutice.archives-ouvertes.fr/edutice-00564559/file/index.html.

Hommage à Jean-Paul Laumond

Marie-Paule Cani [1] et Julien Pettré [2]

Le 20 décembre 2021, nous avons été bouleversés par le décès de Jean-Paul Laumond, pionnier français de la robotique et membre de l'Académie des sciences dont il avait co-présidé avec brio le colloque « Mythes et Machines [3] » le 24 novembre dernier. Le thème de ce tout dernier événement est à l'exacte image de Jean-Paul, entre savoir et culture. Il préférait ainsi évoquer le robot comme une simple « machine », pour rendre limpide le fait que ce dernier n'a d'autre intelligence que celle de ses créateurs.

D'abord professeur de mathématiques, Jean-Paul se réoriente vers le domaine académique et obtient un doctorat de robotique en 1984. Il entre au CNRS en 1985 et effectue l'essentiel de sa carrière au LAAS à Toulouse. Ses recherches portent sur la planification et le contrôle du mouvement des machines autonomes : robots mobiles et humanoïdes. De 2001 à 2003, il crée et dirige la société Kineo CAM qui commercialise les technologies de planification de trajectoire dans le domaine du prototypage virtuel. À l'issue de cette expérience, il se lance sur le nouveau thème de la robotique humanoïde, et codirige le laboratoire franco-japonais JRL de 2005 à 2008. En parallèle, il crée en 2006 l'équipe de recherche Gepetto, dédiée à l'exploration des fondements calculatoires de l'action anthropomorphe. Cette équipe,

1. Professeure à l'École polytechnique.
2. Directeur de recherche Inria.
3. https://www.academie-sciences.fr/fr/Colloques-conferences-et-debats/mythes-et-machines.html.

DOI:10.48556/SIF.1024.19.49

financée en partie par son ERC *advanced grant* (projet Actanthrope, 2014-2018), compte actuellement une trentaine de membres, et dispose d'une plateforme de recherche unique en France avec les deux robots humanoïdes HRP2 et Pyrène. Il rejoint l'équipe parisienne Willow CNRS-ENS-Inria en 2019. Ses travaux sont récompensés par de nombreuses distinctions : IEEE Fellow (2007), prix international IEEE *Inaba Technical Award for Innovation Leading to Production* (2016), élection à l'Académie des technologies (2016), puis à l'Académie des sciences (2017).

Jean-Paul était un chercheur très enthousiaste, qui avait à cœur de partager sa passion pour son domaine. Son passage au Collège de France en 2011-2012, sur la chaire Innovation technologique Liliane Bettencourt, lui a permis d'y créer le premier cours de sa discipline ouvert à tous, introduit par une leçon inaugurale passionnante : « La robotique : une récidive d'Héphaïstos ». Certaines de ses contributions scientifiques ont essaimé, au-delà de la robotique, à l'informatique graphique, pour la planification et la synthèse du mouvement des personnages dans les mondes virtuels. Ses cours étaient des allers-retours permanents entre les questions essentielles qui se posent en robotique pour mettre une machine en action, les algorithmes qui en découlent, et les mathématiques sur lesquelles on pouvait les faire reposer, non seulement pour résoudre les problèmes, mais aussi pour les conceptualiser et les comprendre.

Jean-Paul était chaleureux et ouvert, unanimement apprécié par ses collègues, collaborateurs et étudiants. Indubitablement, cette attitude d'ouverture et d'échange faisait de lui un adepte de la discussion scientifique. Il était un admirateur des problèmes bien posés et de ces questions que la robotique amène, parfois superficielles en apparence mais dont la profondeur conduit à une vision interdisciplinaire : « Comment mettre en équation une trajectoire valide pour un robot posé sur des roues ? Et s'il tire une remorque ? ». En particulier, si la robotique repose inévitablement sur les mathématiques appliquées et sur l'informatique, il aimait à prouver qu'elle pouvait aussi nourrir les mathématiques pures de nouvelles questions théoriques.

Interrogé sur son appartenance à l'informatique, Jean-Paul se disait « ravi d'être du club », mais il répondait également : « *Je préfère me définir comme roboticien, car ce qui m'intéresse c'est le rapport que peut entretenir la machine avec le monde physique, pas seulement le traitement de l'information* ». Cette vision inspirera sans nul doute des générations de chercheurs.

Hommage à Jérôme Monnot

Laurent Gourvès [1]

Repoussée en raison de la crise sanitaire, une conférence internationale en hommage à Jérôme Monnot s'est tenue le 6 décembre 2021 à l'université Paris-Dauphine [2]. *De plus, la revue* Theoretical Computer Science *vient de mettre en ligne un numéro spécial en sa mémoire* [3].

Jérôme Monnot s'en est allé le 11 décembre 2019, à l'âge de 49 ans. Chercheur au CNRS, affecté au laboratoire LAMSADE, Jérôme a communiqué toute sa joie de vivre durant trois décennies à l'université Paris-Dauphine.

Arrivé en 1989 pour entamer un DEUG MASS, il poursuit ses études tout d'abord en génie mathématiques & informatique (licence et maîtrise), puis en méthodes scientifiques de gestion pour son DEA. C'est sous la direction de Vangelis Paschos qu'il soutient en 1998 sa thèse de doctorat en informatique intitulée *Familles d'instances critiques et approximation polynomiale*.

Après trois années passées en qualité de chercheur associé au LAMSADE, il poursuit dans ce même laboratoire son parcours professionnel comme ingénieur de recherche en 2001. Sa persévérance lui permet de décrocher en 2002 un poste de

1. Université Paris-Dauphine, université PSL, CNRS, LAMSADE, Paris.

2. https://jm2021.sciencesconf.org.

3. https://www.sciencedirect.com/journal/theoretical-computer-science/special-issue/10B77W9TTTB.

DOI:10.48556/SIF.1024.19.51

chargé de recherche au CNRS. Il soutient très tôt son habilitation à diriger des recherches (2003), toujours sur le thème de l'approximation polynomiale. En parallèle, Jérôme participe très activement au groupe de travail *Algorithmes à garantie de performance* (AGAPE) du GDR RO où il développe ses premières connexions dans l'Hexagone (CNAM, Evry, LIP6, LIPN...).

Passionné par son métier de chercheur, Jérôme a produit une foule de résultats dans des domaines aussi variés que la recherche opérationnelle, l'optimisation combinatoire, l'approximation polynomiale, la complexité et la théorie des graphes. Au gré des projets ANR auxquels il participe, ses thèmes se diversifient davantage pour s'ouvrir à la théorie des jeux algorithmique (ANR COCA), à l'algorithmique et la complexité paramétrées (ANR TODO), à l'optimisation multi-critère (ANR GUEPARD), ainsi qu'au choix social computationnel (ANR COCORICO). C'est en 2012 que Jérôme est promu directeur de recherche au CNRS. Il avouera par la suite avoir atteint son but professionnel.

Atteint de myopathie, Jérôme a dû composer avec le handicap depuis son plus jeune âge. Voyager et enseigner étaient pour lui quasi impossibles, lui fermant de nombreuses portes pour construire sa carrière académique. Néanmoins, Jérôme a patiemment combattu l'isolement en développant de nombreuses collaborations à la fois locales et à l'étranger (Rome La Sapienza, l'EPFL, l'université de Tel Aviv, la *Business School* d'Athènes, l'université de Twente...) comme en atteste sa longue liste de co-auteurs.

Ceux qui l'ont côtoyé ont souvent loué ses qualités scientifiques, notamment sa capacité à concevoir des réductions complexes sans l'aide d'un papier et d'un crayon, sa curiosité, ainsi que sa connaissance fine de son domaine de recherche. Au laboratoire, il multipliait séances de *coworking* devant un tableau blanc et écriture laborieuse en LaTeX avec sa souris si particulière et son clavier virtuel. C'est dans le partage que Jérôme a construit sa réussite professionnelle : tout nouveau problème, tout encadrement d'étudiant, toute collaboration avec un chercheur de passage au LAMSADE étaient pour lui l'occasion d'inclure d'autres collègues, en particulier les plus jeunes [4].

Jérôme avait une réelle appétence pour l'encadrement et a notamment dirigé ou co-dirigé la thèse de quatre doctorants (Sophie Toulouse, Lydia Tlilane, Mehdi Khosravian Ghadikolaei et Nikolaos Melissinos). De plus, nombreux sont les thésards du LAMSADE qui ont pu bénéficier de son aide pour trouver des pistes afin de débloquer un problème ou pour relire leur mémoire.

Jérôme était aussi très sensible à la répartition équitable des ressources et à la bonne circulation de l'information au sein du laboratoire. C'est selon ces principes qu'il a animé le pôle *optimisation combinatoire, algorithmique* du LAMSADE, et aussi pris part activement au conseil scientifique de l'université Paris-Dauphine.

4. Qui est mieux placé que moi pour le dire ? !

Sur un plan plus personnel, Jérôme avait un humour plein d'autodérision, un détachement vis-à-vis de la maladie qui forçait l'admiration, et une joie de vivre très communicative. Le LAMSADE était sa deuxième maison où résonnaient, dès 14h00, le bruit de son fauteuil électrique et les rires de Corinne, son assistante. Dans son bureau, Jérôme organisait régulièrement des « goûters » où, dans une atmosphère, tantôt studieuse, tantôt festive, se mêlaient chercheurs, administratifs, étudiants, sans distinction d'âge et de discipline.

Ceux qui comme moi ont eu la chance de côtoyer cet amoureux de cinéma et du jeu d'échecs restent à mi-chemin entre le bonheur d'avoir croisé un personnage marquant et la tristesse de l'avoir perdu. Outre ses nombreuses publications, l'héritage de Jérôme Monnot est sa façon d'être et de faire de la recherche, dans le partage, l'entraide et la bonne humeur.

Michel Ugon : « capitaine d'innovation »

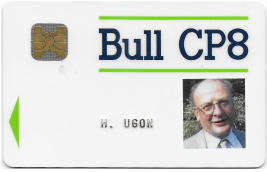
Pierre Paradinas [1]

Rendre hommage aux personnalités qui ont fait progresser l'informatique participe à la construction de la communauté informatique. Parfois, l'hommage est aussi un moyen de rendre sa juste place à un personnage, comme c'est le cas ici pour Michel Ugon. Il fut l'un des acteurs les plus importants à l'origine de la carte à puce mais le nom de Michel Ugon est peu connu en dehors du cercle des spécialistes de cette technologie ayant exercé avant la fin des années 1990. On aurait pu espérer un hommage de la part des grands industriels du domaine qui se sont construits en utilisant ses inventions et ses brevets, mais même les grands journaux d'information semblent avoir oublié son apport.

La carte à puce doit énormément à Michel Ugon. Cette technologie à la croisée de l'informatique, de l'électronique, du *micro-packaging* et de l'impression sécurisée s'est développée initialement dans la deuxième moitié des années 70. L'électronique commençait à se diffuser largement, l'informatique était encore centralisée, le micro-ordinateur apparaissait dans ses marges, et les télécommunications numériques relevaient encore essentiellement de la R&D. Dans ce contexte, cherchant les moyens d'améliorer la sécurité des cartes bancaires à pistes magnétiques, de nombreux travaux s'efforçaient d'inclure une puce électronique dans les 0,76 mm d'épaisseur du support plastique. Les premières puces embarquèrent des fonctions de simple mémorisation avec plus ou moins de sécurité.

1. Professeur au Conservatoire national des arts et métiers (CNAM). L'auteur remercie Pierre-Éric Mounier Kuhn pour la relecture et la famille de Michel Ugon pour les images fournies.

DOI:10.48556/SIF.1024.19.55



La grande innovation apportée par Michel Ugon consista à coupler ces mémoires avec un microprocesseur exécutant des fonctions de gestion et de protection des informations. Le MAM, microcontrôleur auto-programmable monolithique (SPOM en anglais), est au cœur de cette invention, qui sera complétée par d'autres innovations, à la fois pour développer des fonctionnalités du MAM ainsi que pour l'industrialisation de la carte dans des conditions de qualité et d'économie permettant d'en fabriquer des millions à bas coût.

La raison d'être de la carte à puce est de mémoriser des informations et de les protéger. La protection permet de l'utiliser pour identifier une personne (le porteur de la carte) et de contrôler ainsi l'accès à des services pour cette personne (paiement, télécommunications, identité, santé...). La carte contient des programmes qui prennent en charge la gestion de ses données et l'accès aux services, qui mettent en œuvre des fonctions cryptographiques pour sécuriser les données et leurs accès, et qui offrent des fonctions d'authentification, de chiffrement, de signature électronique ou de non-répudiation. À la fin des années 1990, la technologie Java Card s'introduisit dans les cartes pour permettre l'adaptabilité des fonctions des cartes aux besoins de leur promoteur, en permettant de charger au cours du cycle de vie de la carte des programmes et des données — sans remettre en cause la sécurité — avec l'implémentation de machine virtuelle dédiée au microcontrôleur disponible à cette époque.

On trouve les cartes à puces un peu partout aujourd'hui : il s'en fabrique plus de 10 milliards par an ! La forme matérielle de la carte a parfois disparu, les puces se retrouvant dans d'autres dispositifs, mais proposant toujours cette caractéristique forte du couple matériel (microcontrôleur) et logiciel associé, intimement lié avec un haut niveau de sécurité intrinsèque. Cette technologie est portée par des industriels français comme Thales Digital Identity et Idemia, ainsi que Ingenico pour les terminaux.

Michel Ugon, après avoir terminé brillament des études d'ingénieur (major de promo) en 1964 à l'ESEO (École supérieure d'électronique de l'ouest), travailla chez Sexta comme chef du laboratoire, puis chez Jules Richard (instrumentation scientifique) comme chef du laboratoire d'électronique. Il rejoignit en 1971 la Compagnie internationale pour l'informatique (CII, le « champion national » du plan Calcul), d'abord dans la division des périphériques magnétiques, qui produisait des disques

amovibles, puis dans l'équipe développant l'ordinateur Unidata 7740. Après l'absorption de CII par Honeywell-Bull, formant CII-HB en 1976, Michel Ugon s'investit à fond dans le projet carte à puce [2] et déposera de nombreux brevets. La CII-HB filialisera cette activité dans la société Bull-CP8, dont Michel Ugon sera le directeur de la R&D et le DGA. Au début des années 90, quand on évoquait la carte à puce, on la désignait par l'expression « carte CP8 », la marque étant devenu le nom du produit.

Durant cette période, avec deux autres chercheurs — Louis Guillou qui travaille au CCETT et Jean-Jacques Quisquater — Michel Ugon écrit le chapitre *The Smart Card : A Standardized Security Device Dedicated to Public Cryptology* du livre *Contemporary Cryptology : The Science of Information Integrity* de Gus Simmons [3], article qui constituait alors l'une des publications de base sur les cartes à puce. Une autre contribution importante de Michel Ugon a été de prendre en compte la normalisation [4] et d'engager ses forces et celles de Bull CP8 dans ce combat, où les autres industriels n'étaient pas toujours des alliés objectifs. En sa qualité d'expert en sécurité des cartes bancaires, il fut un membre actif de la commission de normalisation ISO 7816, laquelle décrit les caractéristiques physiques de la carte bancaire : dimensions et position des contacts, protocoles de communication et format des données échangées.

Reconnu par ses pairs de l'industrie, il sera le premier président d'Eurosmart, l'association créée il y a 25 ans par les industriels européens de la carte à puce [5], et y présidera le groupe de sécurité ; une question clé dans cette industrie. En 1986, il reçut l'ordre du mérite pour l'invention du MAM.

Michel Ugon était aussi reconnu pour son expertise des cadrans solaires au sein de la Société astronomique de France. Si à la fin du XX^{e} siècle en France on parlait beaucoup des « capitaines d'industrie » pour saluer les grands industriels, Michel Ugon fut un véritable *capitaine d'innovation*.

Pour aller plus loin sur la carte à puce, retrouvez les vidéos de conférences organisées lors de l'exposition sur la carte à puce au Musée des Arts et Métiers :

— « Carte à puce et cryptographie, je t'aime moi non plus », avec Jean-Louis Desvignes et Jean-Jacques Quisquater, `https://mediaserver.cnam.fr/permalink/v125f593a995aqehf7o1/` ;

— « Carte à puce, une histoire de norme et de logiciel », avec René Lozach et Pierre Paradinas, `https://mediaserver.cnam.fr/permalink/v125f593af87b79l0ehz/` ;

— « La carte à puce, de l'invention à l'industrie » avec François Grieu, Philippe Maes, Pierre Mounier Kuhn et Pierre Paradinas, `https://www.arts-et-metiers.net/musee/la-carte-puce-de-linvention-lindustrie`.

2. `https://fr.wikipedia.org/wiki/Carte_%C3%A0_puce`.

3. `https://ieeexplore.ieee.org/book/5265879`.

4. L'ensemble du corpus de normes et standards actuel est assez large, mais peut se retrouver sur les sites de l'IEC/ISO, ETSI, EMVco, *Global Platform*, Oracle (Java Card)...

5. *Smart Card*.

L'écoconception d'un service numérique : des actions pour réduire l'impact environnemental du numérique

Cyrille Bonamy, Cédric Boudinet, Laurent Bourgès, Karin Dassas, Laurent Lefèvre, Benjamin Ninassi, Francis Vivat [1]

Le terme écoconception, surtout appliqué à un service numérique, pourrait apparaître comme une expression un peu vague frôlant le *green washing*. L'objet de cet article est de démontrer qu'il recouvre en réalité des actions concrètes et utiles, dans un monde où l'impact environnemental du numérique est de plus en plus important.

EcoInfo [1], un collectif qui comprend des ingénieurs, des chercheurs et des enseignants-chercheurs des secteurs de la recherche et de l'enseignement supérieur, s'intéresse à cette problématique depuis plus d'une décennie. Le nombre de *think tanks*, d'initiatives publiques et privées, autour de cette problématique n'a cessé de croître ces dernières années [1-8] . Et pour cause, il est réellement urgent d'agir.

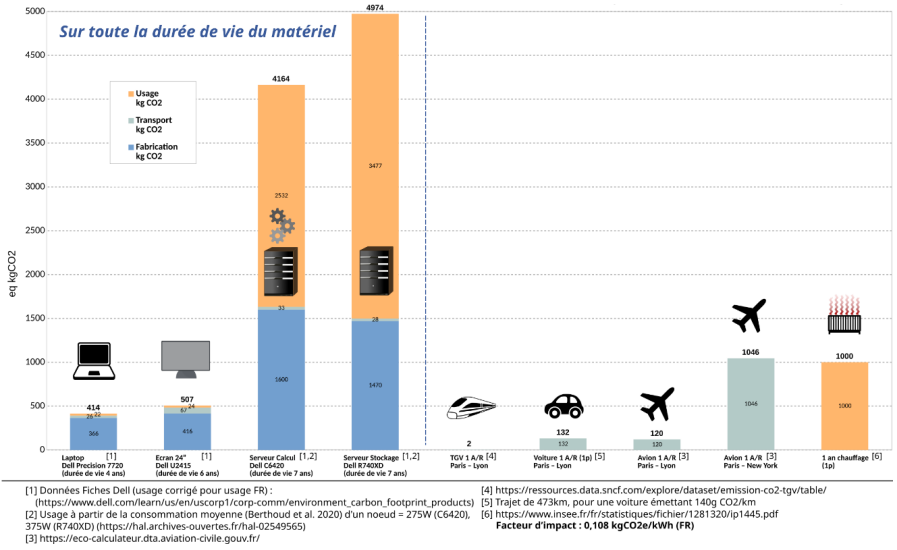
L'écosystème numérique mondial serait à l'origine de 2 % à 4 % — selon les études — des émissions de gaz à effet de serre (GES) sur la planète, soit jusqu'à deux fois plus que le transport aérien. Rien qu'en France, un rapport du Sénat [9] donne 15 millions de tonnes équivalent dioxyde de carbone (CO_2) par an, soit 2% du total des émissions dans l'Hexagone en 2019. Toutes les études s'accordent à dire que cette contribution est en croissance continue [10, 11].

Mais l'impact du numérique ne se limite pas à sa déclinaison la plus médiatisée, les émissions de gaz à effet de serre et le réchauffement climatique.

L'impact majeur est dû à la fabrication du matériel numérique et au traitement des déchets : pollution des sols et des nappes, diminution des ressources en eau, et même

1. Membres du Groupement de service CNRS EcoInfo, https://ecoinfo.cnrs.fr.

DOI:10.48556/SIF.1024.19.59



[1] Données Fiches Dell (usage corrigé pour usage FR) : (https://www.dell.com/learn/us/en/uscorp1/corp-comm/environment_carbon_footprint_products)
[2] Usage à partir de la consommation moyenne (Berthoud et al. 2020) d'un noeud = 275W (C6420), 375W (R740XD) (https://hal.archives-ouvertes.fr/hal-02549565)
[3] https://eco-calculateur.dta.aviation-civile.gouv.fr/
[4] https://ressources.data.sncf.com/explore/dataset/emission-co2-tgv/table/
[5] Trajet de 473km, pour une voiture émettant 140g CO2/km
[6] https://www.insee.fr/fr/statistiques/fichier/1281320/ip1445.pdf
Facteur d'impact : 0,108 kgCO2e/kWh (FR)

FIGURE 1. Comparatif sur les émissions de CO2 (par Jérémy Wambecke et Carole Plasson, Laurent Bourgès).

impact sur la santé et sur la biodiversité. La priorité de l'écoconception de service numérique doit donc être de limiter l'obsolescence des équipements et la nécessité d'en fabriquer de nouveaux.

Pour mieux cerner le périmètre des actions possibles, rappelons d'abord ce qu'est un service numérique, quel est son cycle de vie, et à quelles étapes il est possible de réduire son impact environnemental.

La norme sectorielle ITU L.1410 [12] présente les services numériques comme l'utilisation d'équipements numériques ou de réseaux de télécommunication pour fournir de la valeur à un ou plusieurs utilisateurs.

Nous définissons plus précisément le service numérique comme un ensemble d'information (les données), de traitements (algorithmes, filtrages, simulations), des échanges d'informations, et des interfaces utilisateurs. Un service numérique repose donc sur des infrastructures logicielles (applications, outils, bibliothèques, protocoles), des infrastructures matérielles (serveurs, équipements réseau, terminaux, capteurs), et des personnes (développeurs, administrateurs systèmes et réseaux, chefs de projet, chercheurs, maîtres d'ouvrages et utilisateurs).

Le cycle de vie du service numérique est en général présenté comme sur la figure 2, avec une première phase de définition, collecte des besoins, analyse (phase « avant ») suivie par la conception, le développement logiciel, l'intégration, les tests,

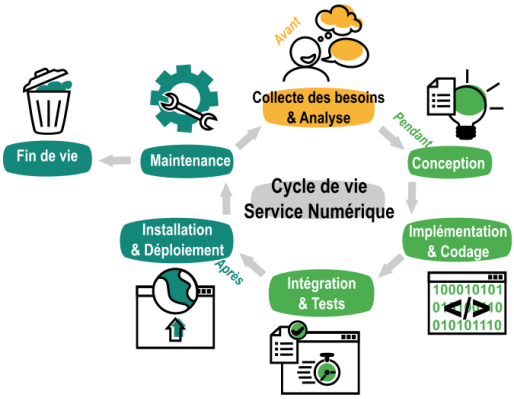


FIGURE 2. Cycle de vie d'un service numérique, Marie Chevallier, 2021.

la mise en production (phase « pendant »), puis le déploiement, l'utilisation et la maintenance (phase « après »), avec une sortie du cycle, la fin de vie. Le cycle peut être parcouru plusieurs fois, dans une démarche d'amélioration continue.

Actions possibles à toutes les étapes du cycle de vie

L'incidence du service numérique sur l'environnement peut se faire jour à toutes les étapes du cycle de vie du service ! De façon générale, plus la prise en compte des aspects environnementaux intervient tôt dans le cycle de vie, plus l'effet est important. Il est donc possible, comme nous allons le voir, d'agir à chaque étape du cycle de vie du service numérique.

Nous présentons ici les grandes lignes des actions possibles, mais nous vous invitons à consulter la plaquette sur les bonnes pratiques en écoconception à destination des développeurs de logiciel [13], rédigée par les auteurs de cet article, et le learning lab d'Inria [14] pour trouver plus de références vers des outils et des méthodologies. Un grand principe d'écoconception, quel que soit le domaine, est d'appliquer le principe « éviter, réduire, compenser » (ERC), notamment les deux premiers points :

— éviter, quand c'est possible, de créer une nouvelle source impactant négativement l'environnement est toujours le meilleur choix ;

— réduire au maximum l'impact quand il est inévitable en faisant preuve de sobriété.

Phase « avant »

Dès la phase « avant », le principe précédent peut s'appliquer. Au début de tout projet de service numérique, on commence par définir l'objet à créer. Quels besoins cherche-t-on à satisfaire ? Quels seront les utilisateurs ? Quels usages en feront-ils ? Pour résumer, dans la phase « avant », on définit le « quoi ». Obtenir des réponses précises et pertinentes à ces questions est une des clefs de la réussite d'un projet, indépendamment de la réalisation technique.

Ainsi, lors de la définition des besoins, il faut éviter le service numérique en trop. Le service qui pollue le moins est celui qui n'existe pas.

Avant de se lancer dans un projet de création, il faut s'assurer que le besoin auquel l'on souhaite répondre n'est pas déjà couvert. D'ailleurs, ce besoin s'inscrit-il dans les objectifs de développement durable [15] ? Ce service aura-t-il un impact positif sur l'environnement, participant ainsi à la compensation de ses propres aspects négatifs ? Se poser la question de l'utilité est la première étape de l'écoconception. Mais une fois le besoin du service numérique justifié, il faut essayer d'éviter l'« obésiciel » en se concentrant sur les fonctionnalités essentielles.

La tentation est grande lors de cette phase de vouloir couvrir un nombre conséquent de besoins, de vouloir toucher le public le plus large possible, et de demander des fonctionnalités pour l'unique raison qu'elles existent dans d'autres services numériques. Ce type de comportement amène à la création de véritables usines à gaz peu utilisables. Plusieurs études [16] révèlent qu'une grande majorité des fonctionnalités des logiciels ne sont que rarement, voire jamais, utilisées (entre 50 % et 80 %). Elles auront pourtant eu un impact environnemental lors de leur production, et leur présence latente dans les logiciels a également un impact résiduel.

Dans un processus cyclique, il est également important de se demander, à cette étape, si les fonctionnalités existantes sont pertinentes, afin de supprimer celles qui sont devenues obsolètes.

Phase « pendant »

Cette phase de réalisation peut elle-même se découper en plusieurs parties : la conception du logiciel, la réalisation, l'intégration et les tests. Pour chacune de ces parties, des actions sont possibles pour tenter de réduire l'impact négatif du service numérique.

Phase « conception »

Concevoir des interfaces sobres, penser à la compatibilité à long terme, réutiliser tout ou partie de logiciels existants, privilégier la basse technologie (ou *low-tech*), planifier la gestion des données (*Data Management Plan* ou DMP) ainsi que la gestion du logiciel (*Software Management Plan* ou SMP) sont autant de leviers à la disposition du développeur lors de la conception du service. C'est aussi le moment

pour le développeur de s'interroger sur les moyens de rendre ses données plus durables, afin d'éviter les développements redondants. Suivre le principe FAIR (*Findable Accessible Interoperable Reusable*), pour les données, comme pour le logiciel d'ailleurs est aussi utile d'un point de vue environnemental.

Phase « réalisation »

La réalisation est le moment où l'on code, où l'on met en œuvre les technologies nécessaires. Cette phase coïncide également avec la réalisation des visuels, icônes, images mais également avec la création des contenus, la collecte et la mise en forme des données, qui devront être accessibles via le service. Au début de cette étape, le choix d'un langage n'est pas anodin d'un point de vue environnemental [13]. Tous les langages et toutes les technologies ne se valent pas, certains peuvent être moins énergivores que d'autres. Ceci dit, attention, les compétences des développeurs et les choix d'architecture ont plus d'impact que le choix du langage : un très bon développeur Python optimisera mieux son code qu'un mauvais développeur C++, ce qui amènera au final à une plus grande réduction de l'impact que le langage lui-même. De même, la réutilisation de code fonctionnel et déjà validé doit être privilégiée, quel que soit le langage. Le contexte, humain ou technique, doit ainsi être un critère déterminant dans le choix des technologies à mettre en œuvre.

On peut essayer aussi de choisir des technologies pérennes, dans la mesure du possible. Les technologies mises en œuvre dans les services numériques évoluent très vite : des nouvelles apparaissent, et certaines disparaissent faute d'une communauté forte pour les maintenir. S'intéresser à l'ancienneté et à l'activité de sa communauté permet d'identifier les technologies pérennes. C'est un critère important à prendre en compte pour éviter de devoir réécrire complètement le code d'un service numérique à cause d'une de ses briques technologiques devenue obsolète.

Un autre principe classique de l'écoconception : éviter la complexité. Un code simple à comprendre et à mettre en œuvre sera aussi plus durable et plus facile à faire évoluer. C'est la base du principe KISS (*Keep It Simple and Stupid*).

Une autre règle d'évitement dans la partie « réalisation » : il est toujours bon d'éviter de réinventer la roue. Dans la mesure du possible, choisissez une licence *open source* et contribuez ainsi au monde du logiciel libre [17, 18]. Non seulement vous permettrez à d'autres de réduire leur propre impact, mais vous pourriez être agréablement surpris par les apports de la communauté à votre création !

Et enfin, on peut mesurer pour mieux réduire. Utilisez des outils de mesure pour évaluer les performances de votre code. Que ce soit la quantité de données échangées, le temps de chargement des pages, le temps d'exécution, le nombre de requêtes... tout est matière à optimisation. L'écoconception est une démarche d'amélioration continue sans fin. La mesure est indispensable à cette démarche afin de savoir quoi et comment améliorer, et d'éviter la tentation de la fonctionnalité en plus

ou l'effet rebond. Néanmoins, trouver le bon compromis entre simplicité et optimisation est souvent un facteur clef.

Attention aux effets rebonds : l'optimisation ne doit pas être un prétexte à l'ajout de nouvelles fonctionnalités non-essentielles, ou à des tests superflus. Optimiser un logiciel peut induire à lancer davantage d'opérations ou traiter davantage de données, donc l'empreinte écologique du service ne sera pas réduite (paradoxe de Jevons). L'optimisation devrait servir simplement à réduire la consommation énergétique et des ressources, et si possible d'arriver plus vite au résultat. Chaque exécution a un impact ! Il est primordial de n'optimiser que ce qui a le plus d'impact (loi de Pareto).

Phase d'intégration et de tests

Dernière étape de la phase « pendant » du cycle de vie, cette phase n'est pas en reste. Il s'agit de mettre la qualité au service de la durabilité. La mise en place de tests permet de détecter les anomalies et d'assurer un important niveau de qualité du service. Ceci va *éviter* les régressions lors de ses évolutions et améliorer la satisfaction des utilisateurs. Si votre service numérique contient trop de *bugs* ou d'instabilités, il risque de ne pas être utilisé ou d'être difficile à maintenir dans le temps : tous les impacts environnementaux engendrés par sa production auront été générés pour rien. Néanmoins, focalisez-vous sur les tests les plus essentiels et méfiez-vous de leur automatisation qui est elle-même une source de gras numérique.

Phase « après »

Une fois le service numérique réalisé et testé, il est encore possible de réduire son impact dans la phase « après », et là encore, à de nombreux niveaux. Lors de la mise en production, choisir un bon hébergeur, dimensionner au plus près du besoin, pratiquer l'amélioration continue et nettoyer les données inutiles, toutes ces actions ont un impact.

Comment distinguer un bon hébergeur d'un mauvais hébergeur ? Pas si simple en fait. Mais il est certain que tous les *datacenters* ne se valent pas en termes de performances environnementales. Choisir un *datacenter* labellisé *Code of Conduct* [19], c'est réduire son impact. D'autres éléments comme sa situation géographique ainsi que son indicateur d'efficacité énergétique (PUE) [20] sont également à prendre en compte.

Choisissez, si possible, une infrastructure (combien de serveurs, de quel type, etc.) au plus proche du besoin réel : vouloir trop anticiper les pics de charge va conduire à un gaspillage de ressources. Privilégiez la mise en œuvre de solutions capables d'allumer et d'éteindre à la demande les ressources nécessaires, et ne laissez pas allumés des serveurs inutilisés.

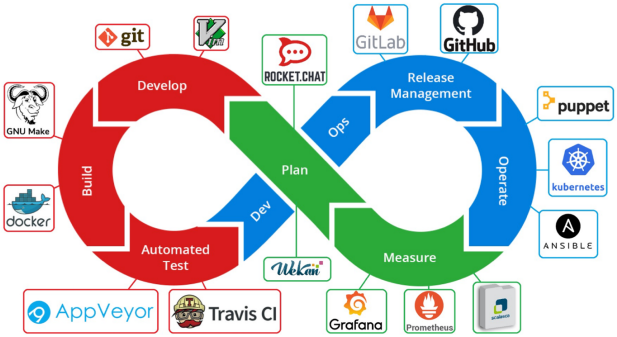


FIGURE 3. Exemples d'outils utilisés pour l'amélioration continue du service numérique (source PNGEgg, adaptée par C. Bonamy).

Une fois le service en production, il peut être utile d'exploiter la supervision et les alertes pour observer les pics CPU, ressources utilisées (disque, réseau), consommation électrique afin de modifier le service numérique pour l'adapter en fonction de l'usage observé (amélioration continue).

Au cours de la vie d'un service numérique, il est commun de voir certaines des données qu'il contient devenir obsolètes. Qu'il s'agisse d'anciens contenus périmés ou de données concernant d'anciens utilisateurs, il est important de régulièrement faire le tri entre les données utiles et celles qui ne le sont plus, de manière à réduire l'impact du stockage.

Phase « fin de vie »

Puis, arrive la fin de vie ; eh oui, pour le service numérique aussi. Même si un service numérique n'est plus utilisé, il va continuer à avoir un impact résiduel sur l'environnement (à cause du stockage de données, de son hébergement et même des accès réseaux réguliers à d'autres services numériques nécessaires à son fonctionnement). Il est donc important de penser à traiter sa fin de vie. Pour ne pas perdre la connaissance produite, pensez à référencer son code dans la plateforme de sauvegarde du patrimoine numérique Software Heritage [21].

Nous avons parcouru ensemble le cycle de vie du service numérique en proposant des actions en fonction des étapes. Mais il existe aussi des éléments auxquels on peut réfléchir à chacune des étapes : la sécurité et l'impact humain.

La sécurité comme facteur de résilience

De l'indisponibilité de votre service jusqu'au vol de données personnelles de vos utilisateurs en passant par l'installation de logiciels malveillants sur votre infrastructure, les conséquences d'une faille de sécurité peuvent être nombreuses. Les actions

correctives, quand elles sont possibles, sont très coûteuses en temps et en énergie, et ont donc un impact sur l'environnement. Le gaspillage est encore plus important si le service doit fermer suite à une attaque. La sécurité doit être une préoccupation présente à toutes les étapes du cycle de vie d'un service numérique, et en lister les bonnes pratiques nécessiterait une fiche concept à part entière. On peut néanmoins citer comme référence les recommandations de l'Agence nationale de la sécurité des systèmes d'information [22] et cet exemple de Cloudflare sur le traitement des *bots* malveillants [23].

L'impact humain

Produire de manière responsable

Chacune de ces étapes va se répéter pendant toute la durée de vie du service numérique, et à chacune de ces étapes des humains collaborent. À travers leur quotidien, ces humains vont avoir un impact : ils vont, par exemple, utiliser du matériel informatique et se déplacer pour venir à leur bureau ou pour se réunir. Mettre en place une politique environnementale d'entreprise permettant de limiter cet impact est crucial. Par exemple, favoriser le télétravail pour les employés qui n'ont pas d'autre solution que la voiture pour venir au bureau va réduire les émissions carbones liées à leurs déplacements. Mettre en place une politique favorisant la prolongation de la durée de vie des ordinateurs et l'utilisation de matériel reconditionné va réduire l'impact de la consommation d'équipements.

Anticiper la transformation des usages

Par son utilisation, le service numérique peut avoir un impact en provoquant une transformation des comportements. Par exemple, une application de réalité augmentée peut avoir été conçue dans l'objectif de donner envie d'aller randonner et de renouer avec la nature, mais aboutir à une incitation de ses utilisateurs à parcourir plus de kilomètres en voiture pour chasser des chimères virtuelles. Il est important de prendre conscience au plus tôt des comportements potentiellement induits par l'utilisation d'un futur service numérique, afin de réduire au maximum l'impact indésirable.

Conclusion

Il faut conserver à l'égard de l'ensemble des bonnes pratiques décrites ici un esprit critique, une capacité d'introspection et une approche systémique. En effet, malgré toutes les bonnes intentions, il est possible que l'application spécifique d'une de ces bonnes pratiques puisse, en fonction du contexte et dans un périmètre donné, générer des effets rebonds et être contre-productive. Il est important de mettre en place des indicateurs (consommation énergétique des équipements mis en œuvre,

quantité de données transférées, temps de chargement, etc.) et de les utiliser pour piloter les évolutions successives afin de s'assurer que la démarche va bien dans le sens de la réduction de l'impact environnemental. L'écoconception est une démarche d'amélioration continue s'inscrivant dans la durée, et il est tout à fait acceptable d'y aller pas à pas, une bonne pratique après l'autre.

Une déclinaison plus spécifique des propositions décrites dans cet article est disponible pour le calcul scientifique dans la plaquette EcoInfo [13]. La prochaine version présentera aussi une fiche *web services*, une fiche rupture et des détails sur la fin de vie. Enfin, nous rappelons que cette plaquette est un complément aux trois plaquettes de bonnes pratiques [24, 25, 26] liées au développement logiciel proposées par le réseau des acteurs du développement logiciel (DevLOG) au sein de l'enseignement supérieur et de la recherche.

Références

[1] https://ecoinfo.cnrs.fr

[2] https://www.arcep.fr/uploads/tx_gspublication/reseaux-du-futur-empreinte-carbone-numeriquejuillet2019.pdf

[3] https://www.institutparisregion.fr/fileadmin/NewEtudes/Etude_1806/Full_report_ENERNUM_MAY_2019-eng.pdf

[4] https://theshiftproject.org/wp-content/uploads/2020/10/Deployer-la-sobriete-numerique_Rapportcomplet_ShiftProject.pdf[3.5]https://www.greenit.fr/empreinte-environnementale-du-numerique-mondial

[5] https://www.greenit.fr/empreinte-environnementale-du-numerique-mondial

[6] https://institutnr.org/inr-numerique-responsable

[7] https://ecoresponsable.numerique.gouv.fr/a-propos

[8] https://boavizta.org

[9] https://www.senat.fr/rap/r19-555/r19-5550.html#toc0

[10] http://www.senat.fr/rap/r19-555/r19-555-syn.pdf

[11] https://theshiftproject.org/article/pour-une-sobriete-numerique-rapport-shift/

[12] https://www.itu.int/rec/T-REC-L.1410/fr

[13] Plaquette EcoInfo, https://hal.archives-ouvertes.fr/hal-03009741

[14] https://learninglab.gitlabpages.inria.fr/mooc-impacts-num/mooc-impacts-numressources/Partie3/FichesConcept/FC3.4.2-bonnespratiques-MoocImpactNum.html

[15] https://www.un.org/sustainabledevelopment/fr/objectifs-de-developpement-durable/

[16] Suja Thomas, *Feature Adoption Report*, Pendo, 2019, https://go.pendo.io/rs/185-LQW-370/images/2019%20Feature%20Adoption%20Report%20Digital.pdf

[17] Logiciel libre et développement durable, http://ll-dd.ch

[18] https://www.etalab.gouv.fr/accompagnement-logiciels-libres

[19] https://ecoinfo.cnrs.fr/2020/05/19/guide-des-bonnes-pratiques-du-code-de-conduite-europeen-sur-les-datacentres

[20] https://fr.wikipedia.org/wiki/Indicateur_d'efficacité_énergétique

[21] https://www.softwareheritage.org/?lang=fr

[22] Sécuriser un site web, ANSSI, https://www.ssi.gouv.fr/guide/recommandations-pour-la-securisation-des-sites-web

[23] John Graham-Cumming, *Cleaning up bad bots*, Cloudflare, 23/09/2019, https://blog.cloudflare.com/cleaning-up-bad-bots

[24] https://hal.archives-ouvertes.fr/hal-02083801

[25] https://hal.archives-ouvertes.fr/hal-02400300

[26] https://hal.archives-ouvertes.fr/hal-02399517

Raconter la science en temps de crise, discours d'inauguration de la journée Sciences et médias

Yves Sciama et Audrey Mikaëlian [1]

Les journées « Sciences et médias » réunissent régulièrement les acteurs de la science et des médias autour de thématiques concernant la place de la science dans les médias. Cette année 2022, la journée s'est déroulée le 25 janvier dans les locaux de la Bibliothèque nationale de France et portait sur la science en temps de crise. Voici le discours inaugural de cette journée écrit par Yves Sciama et Audrey Mikaëlian ; il trace parfaitement les contours de cette journée, dont les débats passionnants et leur illustration au fil de la tablette graphique par le talentueux Guillaume Monnain sont accessibles sur le site des journées [2].

Merci d'être venus si nombreux participer, dans ce magnifique auditorium ou à distance, à cette journée « Sciences et médias » intitulée « Raconter la science en temps de crise ».

Je dois dire que cette année, le thème s'est imposé de lui-même et a fait l'unanimité du comité éditorial qui a programmé cette journée. Scientifiques, chercheurs, journalistes, médiateurs, communicants... depuis deux ans, nous en sommes tous à raconter la science en temps de crise. Et pourtant, en y réfléchissant de plus près, la question de savoir si l'on est vraiment en crise peut se poser. On pourrait se demander, dans le fond, si le mot *crise* est vraiment le plus juste pour décrire ce que nous vivons...

1. Journalistes scientifiques et membres de l'Association des journalistes scientifiques de la presse d'information (AJSPI) qui co-organise la journée.

2. http://sciencesetmedias.org.

DOI:10.48556/SIF.1024.19.69

Après tout, cela fait déjà 10 ans que la philosophe Myriam Revaud d'Allonnes a écrit « La crise sans fin ». Dans cet ouvrage de 2012, elle plaidait à juste titre qu'une crise, au sens strict, est un moment décisif, un moment au cours duquel un système présumé stable se déstabilise et arrive à une bifurcation — et non pas un état durable... Or il est largement évident que la crise environnementale, dont nous parlerons cette après-midi, n'en est pas une selon cette définition, puisqu'elle a émergé depuis une cinquantaine d'années, et que les bouleversements qui lui sont associés vont perdurer au moins pour des décennies, et même en réalité des millénaires. Sur le plan de l'environnement, nous ne sommes donc pas dans une crise mais dans une phase de transition radicale et durable du monde. Et même en s'en tenant à la seule crise sanitaire, il y a des bonnes raisons de penser qu'il ne s'agit plus d'une crise à proprement parler, mais plutôt d'une nouvelle normalité. Le — ou plutôt les — coronavirus ne vont pas disparaître. Et cette nouvelle normalité sera sans doute faite de vagues d'infections, dont rien ne dit qu'elles ne concerneront que des coronavirus, séparées par des stases, épidémiologiquement plus ou moins longues et calmes. Mais récuser le terme de crise a-t-il vraiment un intérêt ? Car en matière d'information et de savoir scientifique, il saute aux yeux que la période actuelle a de nombreuses caractéristiques que l'on associe à ce terme : la politisation des débats, leur électrisation, leur polarisation, et puis la demande de réponses rapides, et une conflictualité inédite, probablement alimentée par la montée de multiples peurs. Et nous constatons tous, dans le courrier des lecteurs pour nous les journalistes, mais aussi sur les réseaux sociaux, et tout simplement dans notre vie quotidienne, la montée d'une défiance généralisée vis-à-vis des politiques, qui s'est propagée aux médias, et qui devient parfois une réelle hostilité.

Quant à la science, elle n'en est peut-être pas là, comme nous le dira Michel Dubois qui préfère parler de désenchantement. Il n'en reste pas moins que l'on voit désormais apparaître et circuler, y compris dans des milieux sociaux instruits et habituellement éclairés, des informations fausses de plus en plus nombreuses. Et que de prétendus experts, sur la base souvent uniquement d'un diplôme de médecine ou de science, défendent avec aplomb des points de vue à mille lieues de ce que dit la littérature, et trouvent facilement des audiences plus importantes que celles des véritables spécialistes. Il ne manque peut-être pas grand-chose à ce désenchantement de la science pour se transformer en désamour, voire en animosité.

Nous sommes donc bel et bien en crise — simplement, cette crise va durer et cette durée est une raison pressante de poser les questions que nous allons nous poser aujourd'hui. Quelles sont les implications de cette nouvelle donne pour les scientifiques et les journalistes ? Car il nous faut, chacun à notre place, et aussi en interagissant davantage, contribuer à sortir de l'ornière. Une fois l'actuelle phase de sidération passée, il va bien falloir ouvrir le champ des possibles, œuvrer à ce que soient prises des décisions rationnelles et favorables au plus grand nombre, favoriser un débat public éclairé et factuel.

Oui il va falloir chercher un nouvel équilibre dans ce monde si bouleversé. Il va falloir recoller les morceaux de notre société déchirée, et trouver le moyen de transformer cette crise en une plate-forme commune, en terrain d'entente, en point de départ d'une reconstruction. Peut-être que le discours et la connaissance scientifiques peuvent être un moyen de prendre ce chemin ?

Alors, une chose est sûre, chacun doit résister à la tentation de l'abstention et de la passivité. Certes, raconter la science en temps de crise, que l'on soit journaliste ou scientifique, c'est affronter des pièges, des défis et des dangers multiples. C'est s'exposer dans une atmosphère particulièrement électrique, c'est braver l'erreur, mettre en jeu sa réputation, c'est prendre des risques parfois économiques, voire juridiques ou même physiques.

L'un des points communs qu'ont les scientifiques et les journalistes, c'est qu'ils tentent avec leurs outils spécifiques d'approcher au plus près ce que — faute de mieux — on peut appeler la vérité. Et qu'ils sont, de ce point de vue, au service du public. Or la société a besoin de réponses, elle doute, elle s'interroge... Et particulièrement en ce moment, elle a une vraie soif de science. A nous, journalistes, scientifiques, de savoir y répondre, de trouver le ton, les mots, les formes qui conviennent le mieux, à nous d'apprendre des erreurs qui ont été commises, à nous de continuer à progresser. C'est en tous cas l'objectif que nous avons voulu donner à cette journée, et nous sommes certains qu'elle sera fructueuse.

Note du comité d'organisation :

Que le thème de ces journées porte sur les médias traditionnels et numériques, les médias destinés à la jeunesse, la désinformation scientifique, ou bien encore la place des femmes scientifiques dans les médias (c'était le thème en 2020), l'objectif de ces journées est d'abord de dresser un constat des difficultés ou problèmes existants, puis de proposer des solutions. En 2022, à l'heure où science et technologie sont au cœur des enjeux sociétaux (climat, énergie, 5G, biotechnologies, vaccination, etc.), on peut se poser des questions sur la façon dont les médias ont couvert la pandémie de COVID. Comment expliquer la cacophonie délétère qui s'est installée ? Quelle a été la place des journalistes scientifiques dans les rédactions ? Fallait-il laisser le traitement de cette information aux experts, et lesquels ? Pour l'avenir de notre santé comme de notre démocratie, il nous semble indispensable d'améliorer le traitement des questions et des controverses scientifiques.

Bilan du prix de thèse Gilles Kahn 2021

Clémentine Maurice et Charlotte Truchet

Le prix de thèse Gilles Kahn 2021, décerné par la SIF et patronné par l'Académie des sciences, a été attribué à :

GABRIELLE DE MICHELI
pour sa thèse « Discrete Logarithm Cryptanalyses : Number Field Sieve and Lattice Tools for Side-Channel Attacks », soutenue le 25 mai 2021 au LORIA à Nancy sous la direction de Pierrick Gaudry et Cécile Pierrot.

Les accessits (par ordre alphabétique) ont été décernés à :

REBECCA FRIBOURG
pour sa thèse intitulée « Contribution of the study of factors influencing the sense of embodiment towards avatars in virtual reality » soutenue à l'IRISA à Rennes, sous la direction de Anatole Lécuyer, Ferran Argelaguet et Ludovic Hoyet.

THIBAULT GROUEIX
pour sa thèse « Learning 3D Generation and Matching », soutenue à l'Université Paris-Est, sous la direction de Mathieu Aubry et Renaud Marlet.

CHARLIE JACOMME
pour sa thèse « Proofs of security protocols - symbolic methods and powerful attackers », soutenue à l'Université Paris Saclay, sous la direction de Hubert Comon et Steve Kremer.

DOI:10.48556/SIF.1024.19.73

Les thèses des lauréats sont accessibles sur le site de la SIF :

http://www.societe-informatique-de-france.fr/recherche/prix-de-these-gilles-kahn/

et les différents travaux de thèse sont racontés dans la rubrique « Il était une fois... ma thèse » du blog binaire à l'adresse :

https://www.lemonde.fr/blog/binaire/il-etait-une-fois-ma-these/

Un résumé de chacune de ces thèses vous est également proposé dans les articles suivants de ce numéro de 1024.

Le jury a reçu cette année 30 dossiers couvrant un très large spectre de travaux de recherche et en provenance de nombreux laboratoires de recherche en informatique. Quelques thématiques et laboratoires étaient cependant absents. Nous rappelons donc que ce prix s'adresse à l'ensemble de la discipline informatique et à l'ensemble des laboratoires ou centres de recherche français.

Le jury 2021 était présidé par Charlotte Truchet, assistée par Clémentine Maurice, secrétaire du prix. Le jury était constitué de Mathieu Acher, Marie Albenque, Chloé-Agathe Azencott, Raphaëlle Chaine, Thomas Chatain, Caroline Collange, Arnaud de Mesmay, Thomas Debris-Alazard, Thomas Degueule, Omar Fawzi, Jean-Daniel Fekete, Jérôme Féret, Laure Gonnord, Amaury Habrard, Christine Largeron, Cédric Lauradoux, Olivier Ly, Vincent Nicomette, Laurent Perron, Maria Potop-Butucaru, Vivien Quema, Pierre Senellart, Tristan Vaccon.

Merci à toutes les candidates et tous les candidats pour la qualité de leurs travaux et merci aux membres du jury pour leur participation.

Exponentiation modulaire pour la cryptographie à clé publique

Gabrielle De Micheli [1]

Gabrielle De Micheli a soutenu sa thèse [2] le 25 mai 2021 opérée au sein de l'université de Lorraine sous la direction de Pierrick Gaudry et de Cécile Pierrot. Vous trouverez ci-dessous un résumé de cette thèse.

La cryptographie à clé publique

La cryptographie s'intéresse au problème de l'échange de messages chiffrés, c'est-à-dire inintelligibles, que seul un récepteur légitime peut déchiffrer, donc lire. Afin d'assurer une transmission sécurisée de ces messages, une clé secrète est généralement partagée entre l'expéditeur et le destinataire. Cela pose la difficulté importante d'échanger de manière sécurisée la clé secrète mentionnée ci-dessus.

Au début des années 1970, Merkle a commencé à s'écarter de ce concept de clé partagée et ses idées publiées en 1978 [4] ont été reprises dans l'article fondateur de Diffie et Hellman [2], *New directions in Cryptography*. Dans leur article, Diffie et Hellman formalisent la notion de cryptographie à clé publique où deux clés mathématiquement liées sont générées et utilisées : une clé publique et une clé secrète. Un message est ensuite chiffré à l'aide de la clé publique du récepteur. Ce dernier sera alors le seul capable de déchiffrer le message à l'aide de la clé secrète correspondante.

Les cryptosystèmes à clé publique, également connus sous le nom de protocoles asymétriques, sont tous construits en utilisant la notion de fonction trappe. *Cette*

1. gdemicheli@eng.ucsd.edu.

2. https://www.theses.fr/2021LORR0104.

DOI:10.48556/SIF.1024.19.75

terminologie désigne une fonction qui est facile à évaluer pour toute entrée mais dont il est difficile de calculer l'antécédent d'une valeur donnée. Cette notion correspond bien aux exigences d'un protocole asymétrique. En effet, pour qu'un protocole soit sûr et efficace, le déchiffrement d'un message sans la clé secrète doit être proche de l'impossible, alors que le chiffrement d'un message et le déchiffrement avec la clé secrète doivent être faciles, c'est-à-dire réalisés uniquement avec des opérations simples.

C'est naturellement vers des problèmes mathématiques difficiles que les cryptographes se sont tournés pour trouver des primitives appropriées pour leurs protocoles. Historiquement, deux candidats ont émergé : la multiplication de deux nombres premiers et l'exponentiation modulaire. Inverser ces fonctions demande de savoir, respectivement, factoriser un entier ou calculer un logarithme discret. La difficulté de la factorisation est au cœur du cryptosystème RSA, bien connu et déployé [5]. Dans ce travail, nous nous sommes concentré sur le second candidat : l'exponentiation modulaire et son opération inverse, le calcul d'un logarithme discret.

Exponentiation modulaire et logarithme discret

L'exponentiation modulaire consiste à calculer le reste d'une division euclidienne d'un entier g élevé à une puissance x par un entier positif N, en calculant $g^x \pmod N$. Cette opération est fondamentale dans la théorie computationnelle des nombres où elle se retrouve par exemple dans le petit théorème de Fermat utilisé pour le test de primalité. L'exponentiation modulaire est également largement utilisée en cryptographie à clé publique, où les éléments de groupes tels que les groupes multiplicatifs des corps finis, $\mathbb{Z}/N\mathbb{Z}$ ou le groupe des points rationnels des courbes elliptiques, sont souvent élevés à de grandes puissances.

Pour des raisons pratiques, les opérations utilisées dans les protocoles cryptographiques doivent être faciles et efficaces à réaliser. L'attrait de l'exponentiation modulaire pour la cryptographie provient en partie de la simplicité de son calcul. Cependant, cette méthode serait très inefficace dans un contexte cryptographique en raison de la taille des nombres impliqués. Par conséquent, afin de construire des protocoles cryptographiques pratiques qui utilisent l'exponentiation modulaire, l'opération doit être effectuée de manière efficace.

L'efficacité des algorithmes qui calculent l'exponentiation modulaire dépend de divers paramètres tels que le groupe considéré, la représentation de l'exposant ou le matériel utilisé. Comme l'exponentiation modulaire est présente dans de nombreux protocoles, et qu'elle est souvent l'opération la plus coûteuse du protocole, au fil des années, de nombreux algorithmes optimisés se sont accumulés pour améliorer son calcul.

Bien que de nombreux efforts aient été déployés pour optimiser les algorithmes pour l'exponentiation modulaire, ces optimisations ont fait apparaître des vulnérabilités exploitables. L'exécution du code peut alors générer des fuites observables à

partir desquelles des informations peuvent être déduites sur l'exposant. Le caractère spécifique des informations divulguées dépend des détails de la mise en œuvre de l'algorithme et souvent du matériel lui-même. Les attaques par canaux auxiliaires et en particulier les attaques par cache sont les principales menaces à prendre en compte lors de l'utilisation d'un algorithme d'exponentiation modulaire rapide pour un protocole.

L'opération inverse de l'exponentiation modulaire est le calcul d'un logarithme discret. L'étude des logarithmes discrets et des algorithmes associés précède leur utilisation en cryptographie. En effet, dès le 19^{e} siècle, les logarithmes de Zech sont utilisés pour accélérer les opérations arithmétiques dans les corps finis.

En cryptographie, le protocole Diffie-Hellman datant de la fin des années 70 a marqué un tournant dans l'étude des logarithmes discrets. L'utilisation plus récente des logarithmes discrets dans les protocoles basés sur les couplages, qui a débuté au début des années 2000, a relancé l'intérêt pour le sujet. Concrètement, un logarithme discret est défini comme suit.

Définition 1 (Logarithme discret)**.** *Étant donné un groupe cyclique fini G d'ordre n, un générateur $g \in G$ et un élément $h \in G$, le logarithme discret de h en base g est l'élément $x \in [0,n[$ tel que $g^x = h$.*

Cette définition pose le problème suivant.

Définition 2 (Le problème du logarithme discret (DLP))**.** *Étant donné un groupe cyclique fini G d'ordre n, un générateur $g \in G$, et un élément $h \in G$, trouver x tel que $g^x = h$.*

Ce problème est considéré comme difficile pour la plupart des groupes G et constitue donc un candidat prometteur pour la cryptographie à clé publique. Des cryptosystèmes largement déployés, tels que le protocole d'échange de clés Diffie-Hellman, le protocole de chiffrement d'ElGamal ou les protocoles de signature tels que (EC)DSA basent leur sécurité sur des hypothèse liées à la difficulté du problème du logarithme discret. La sécurité des protocoles basés sur les couplages repose également sur la difficulté du problème du logarithme discret.

Contributions

L'objectif de ce travail est de répondre à la question suivante.

Question 1. *Comment évaluer la sécurité des protocoles dans lesquels une exponentiation modulaire impliquant un secret est effectuée ?*

La réponse à cette question se divise en deux points.

(1) La résolution du problème du logarithme discret donne un accès direct à l'exposant, donc au secret. Ainsi, nous voulons estimer la difficulté de DLP dans les groupes utilisés dans les protocoles cryptographiques considérés.

(2) L'étude des vulnérabilités d'implémentation pendant l'exponentiation rapide peut également conduire à découvrir l'exposant secret. Ainsi, nous voulons également évaluer et étudier les attaques rendues possibles grâce aux informations fuitées par des canaux auxiliaires.

Estimation de la difficulté de DLP dans les corps finis

Une façon d'estimer la sécurité des protocoles basés sur la difficulté du problème du logarithme discret est d'étudier directement la complexité des algorithmes qui résolvent ce dernier. Cela dépend fortement du groupe considéré. Dans ce travail, nous nous concentrerons sur l'estimation de la difficulté de DLP pour des corps finis spécifiques, situés à un cas frontière. L'estimation de la difficulté du DLP pour ces corps finis a un impact significatif sur notre compréhension de la sécurité des protocoles largement déployés.

Notre première motivation concerne la sécurité des protocoles basés sur les couplages, des applications bilinéaires non dégénérées. Ces dernières envoient une paire d'éléments, généralement des points sur une courbe elliptique bien définie vers un élément d'un corps finis. Si nous voulons qu'un couplage soit sûr, nous voulons équilibrer la complexité de l'algorithme en racine carrée qui calcule les logarithmes discrets dans le sous-groupe pertinent de la courbe elliptique considérée, et la complexité de l'algorithme le plus rapide qui résout DLP dans le corps fini. Ceci nous a amené à étudier les algorithmes de la famille du calcul d'indice à la frontière entre les corps finis dits de petite caractéristique et ceux de caractéristique moyenne. La complexité asymptotique de ces algorithmes à ce cas frontière était, jusqu'à ce travail, inexistante dans la littérature. Grâce à cette analyse, nous avons finalement pu fournir des informations supplémentaires sur les paramètres de sécurité des protocoles basés sur les couplages.

Une autre façon d'obtenir de meilleures estimations de sécurité consiste à réaliser des expériences à grande échelle avec des variantes de l'algorithme Number Field Sieve [3]. En effet, l'algorithme Number Field Sieve a donné lieu à de nombreuses variantes, chacune tentant de réduire la complexité asymptotique de l'algorithme original. L'une de ces variantes est le *Tower Number Field Sieve* (TNFS) [1]. Ce dernier exploite la structure algébrique des tours de corps de nombres. Malgré le fait qu'en théorie la variante est plus que prometteuse, aucune implémentation et donc aucun calcul record n'avait été fait en utilisant TNFS, jusqu'à ce travail.

Un obstacle majeur à une mise en œuvre efficace de TNFS est la collection de relations algébriques où des équations entre de petits éléments de corps de nombres doivent être trouvées. Le cas de TNFS est plus complexe que celui de NFS car

cette collecte de relations se produit en dimension supérieure à 2. Cela nécessite la construction de nouveaux algorithmes de crible qui restent efficaces lorsque la dimension augmente.

Nous surmontons cette difficulté en considérant un algorithme d'énumération sur les réseaux que nous adaptons à ce contexte spécifique. Cela nous a permis d'effectuer le premier calcul record d'un logarithme discret avec TNFS dans un corps fini de 521 bits $\mathbb{F}_{p^6}$. Le corps fini cible $\mathbb{F}_{p^6}$ choisi est de la même forme que les corps finis utilisés dans les récentes preuves zéro *knowledege* de certaines blockchains.

Notre analyse sur la sécurité des protocoles basée sur les couplages ainsi que le calcul record avec TNFS contribuent à estimer la difficulté du DLP dans les corps finis en étudiant les complexités asymptotiques des algorithmes pertinents et en fournissant un calcul record avec TNFS. Les considérations faites sur la sécurité des couplages devraient compléter les estimations pratiques trouvées dans la littérature et, espérons-le, orienter les cryptanalystes vers les bons choix de paramètres. Les performances pratiques de TNFS avec notre nouvel algorithme de crible sont prometteuses et indiquent que des corps finis plus grands pourraient être atteints en un temps raisonnable. En général, les calculs records fournissent des indications supplémentaires sur l'écart entre les tailles de clés recommandées pour les protocoles basés sur DLP et ce qui est faisable sur le plan informatique.

Exploitation des vulnérabilités de l'exponentiation rapide

La sécurité des protocoles déployés ne dépend pas seulement de la difficulté du problème mathématique sous-jacent, mais aussi de l'implémentation des algorithmes concernés.

Les implémentations vulnérables de l'exponentiation modulaire rapide ont souvent été la cible d'attaques par canaux auxiliaires où des informations secrètes sont récupérées en créant des liens observables entre les différentes unités d'exécution du CPU. En particulier, les attaques temporelles exploitent les variations du temps d'exécution qui sont courantes dans les algorithmes d'exponentiation modulaire.

Nous présentons dans ce travail un aperçu des techniques connues pour récupérer des clés secrètes à partir d'informations partielles ainsi que deux cas pratiques. Nous étudions d'une part la sécurité de l'implémentation d'Intel du protocole EPID (*Extended Privacy ID*), un protocole d'authentification et d'attestation à distance. Nous identifions une faiblesse d'implémentation qui fait fuiter des informations via un canal auxiliaire du cache. Cette fuite d'information nous permet de monter une approche basée sur les réseaux pour résoudre le *Hidden Number Problem*, que nous adaptons à la preuve *zero-knowledge* du protocole EPID, étendant ainsi les attaques antérieures sur les systèmes de signature. Ce travail montre qu'un fournisseur d'attestation malveillant peut utiliser l'information divulguée pour briser les garanties de non-liaison d'EPID.

Nous nous concentrons également sur la sécurité du protocole ECDSA lorsque le nonce k utilisé dans l'algorithme de signature comme exposant modulaire est exprimé sous la forme wNAF. Nous trouvons la clé secrète avec seulement 3 signatures, atteignant ainsi une limite théorique connue, alors que les meilleures méthodes précédentes nécessitaient au moins 4 signatures en pratique. Étant donné un modèle de fuite spécifique, notre attaque est plus efficace que les attaques précédentes et, dans la plupart des cas, a une meilleure probabilité de succès. Nous fournissons également une première analyse de la résistance aux erreurs de EHNP.

En considérant des cibles réelles telles que EPID dans l'architecture d'Intel et l'algorithme ECDSA largement déployé, nous montrons tout au long de ces travaux que même si les bons paramètres sont pris en compte pour que le problème du logarithme discret reste suffisamment difficile à résoudre à des fins cryptographiques, les attaques peuvent provenir d'implémentations vulnérables de l'exponentiation modulaire. Afin d'évaluer réellement la sécurité des protocoles à clé publique déployés, il faut donc considérer simultanément les menaces provenant de la primitive mathématique elle-même et de l'implémentation des algorithmes.

Références

[1] R. Barbulescu, P. Gaudry, and T. Kleinjung. The tower number field sieve. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 31–55. Springer, 2015.

[2] W. Diffie and M. Hellman. New directions in cryptography. *IEEE Trans. Inf. Theor.*, 22(6) :644–654, Sept. 2006.

[3] A. K. Lenstra, H. W. Lenstra, M. S. Manasse, and J. M. Pollard. The number field sieve. In *The development of the number field sieve*, pages 11–42. Springer, 1993.

[4] R. C. Merkle. Secure Communications over Insecure Channels. *Commun. ACM*, 21(4) :294–299, Apr. 1978.

[5] R. L. Rivest, A. Shamir, and L. Adleman. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Commun. ACM*, 21(2) :120–126, Feb. 1978.

Contribution à l'étude des facteurs influençant le sentiment d'incarnation envers un avatar en réalité virtuelle

Rebecca Fribourg [1]

Rebecca Fribourg a soutenu sa thèse [2] le 04 novembre 2020 opérée à l'université Rennes 1 au sein des instituts Inria et Irisa sous la direction d'Anatole Lécuyer, de Ferran Argelaguet et Ludovic Hoyet. Vous trouverez ci-dessous un résumé de cette thèse.

Ma thèse s'inscrit dans le domaine de la réalité virtuelle (RV) et plus précisément des « avatars », c'est-à-dire la représentation numérique des utilisateurs dans les environnements virtuels. Dans ce contexte, les défis consistent à développer de nouvelles approches permettant de prendre en compte à la fois les aspects technologiques (dispositifs, algorithmes, complexités) et cognitifs (psychologie, perception, etc.) pour permettre à l'utilisateur d'exploiter pleinement les nouvelles possibilités proposées par les avatars en réalité virtuelle. Par conséquent, il est aujourd'hui nécessaire de mieux comprendre les processus qui définissent comment les utilisateurs perçoivent leur avatar et interagissent à travers lui. Comprendre et identifier les facteurs qui impactent le sentiment d'incarnation d'un utilisateur envers son avatar est donc un réel challenge, et j'ai choisi de l'explorer dans le cadre de mon doctorat. Dans ce travail, nous avons défini trois axes de recherche pour explorer l'influence de plusieurs facteurs sur le sentiment d'incarnation (SI), en nous basant sur une catégorisation qui

1. rebecca.fribourg@ec-nantes.fr.
2. http://www.theses.fr/2020REN1S052.

DOI:10.48556/SIF.1024.19.81

ne prend pas seulement en compte les facteurs liés à l'avatar, mais aussi les facteurs liés à l'environnement virtuel et à l'utilisateur.

Influence des facteurs liés à l'environnement virtuel

Les environnements virtuels (EVs) peuvent être caractérisés par une multitude de facettes, telles que leur style de rendu ou leur réalisme, leur degré d'interactivité et la quantité de retours sensoriels qu'ils fournissent. Les caractéristiques des EVs sont connues pour influencer les expériences de RV des utilisateurs et, plus précisément, pour influencer le sentiment de présence des utilisateurs, un autre *quale* qui fait référence au « sentiment d'être dans le monde virtuel » [12]. Cependant, l'impact des caractéristiques de l'environnement virtuel sur le SI reste rarement étudié. En particulier, nous avons identifié deux aspects de l'EV susceptibles d'influencer le SI : la dimension sociale de l'EV (c'est-à-dire la présence d'autres utilisateurs partageant le même EV) et l'introduction de menaces envers l'avatar dans l'EV.

Environnements virtuels partagés

De plus en plus d'expériences de RV partagée de haute qualité sont maintenant proposées par les développeurs de RV. Ces configurations permettent à plusieurs utilisateurs d'être immergés dans le même environnement virtuel sans nécessairement être physiquement au même endroit. Ils ont également la possibilité d'interagir simultanément les uns avec les autres et avec l'EV. Ces progrès ont revigoré les intérêts de la recherche dans les EV partagés [1, 10, 13]. Afin d'évaluer l'effet de ces EV partagés sur l'expérience des utilisateurs, le sentiment de présence a fait l'objet d'une étude approfondie. Il a été démontré que le fait de voir d'autres utilisateurs dans l'environnement virtuel pouvait être considéré comme une preuve de sa propre existence dans l'environnement virtuel et pouvait accroître le sentiment de présence. Cependant, alors que le sentiment de présence a été étudié dans les EV partagés, les études du SI semblaient alors se concentrer uniquement sur les expériences à utilisateur unique. Il n'était donc pas encore très bien connu en quoi le partage d'expériences virtuelles avec un autre utilisateur incarné dans un avatar pourrait influencer son propre SI. Nous avons par conséquent décidé d'explorer cette question.

À cette fin, nous avons étudié en premier lieu l'influence des environnements virtuels partagés sur le SI, dans une étude [5] où les utilisateurs accomplissaient ensemble une tâche dans le même environnement virtuel. J'ai conçu un système d'immersion permettant l'interaction collective de deux utilisateurs incarnés chacun dans un avatar et j'ai ensuite évalué ce système en mesurant le SI respectif des utilisateurs envers leur avatar. Grâce à cette étude, il a pu être montré que partager l'environnement virtuel EV avec quelqu'un d'autre ne détériorait pas son propre sentiment d'incarnation SI, ce qui est une bonne chose pour toutes les applications de RV impliquant des avatars.

FIGURE 1. Expérience étudiant l'influence du partage de l'EV sur le SI [5]. Les utilisateurs étaient immergés successivement seuls, seuls en face d'un miroir, puis à deux dans l'EV, avec un ordre variable.

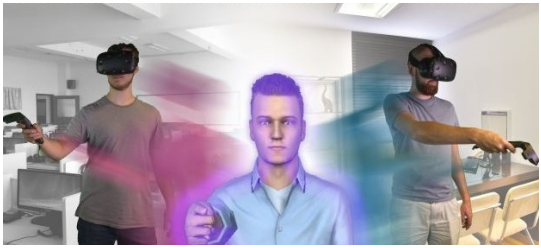


FIGURE 2. Illustration de l'expérience explorant le sentiment de contrôle envers un avatar partagé entre deux utilisateurs [8]. La position et orientation du bras droit de l'avatar correspont à la moyenne pondérée entre la position et orientation des bras des deux utilisateurs, avec un niveau de partage variable.

Dans une autre étude [8] menée avec Nami Ogawa une doctorante de l'université de Tokyo en visite au laboratoire, nous avons exploré plus avant le contexte de l'environnement virtuel partagé en étudiant l'influence du partage d'un avatar virtuel avec un autre utilisateur. Plus précisément, nous nous sommes intéressés au partage du contrôle de l'avatar, et à la manière dont le poids de contrôle partagé (qui était modulé) influencait le sentiment d'agentivité et les actions motrices des utilisateurs. Nous avons montré que deux utilisateurs pouvaient encore ressentir un sentiment de contrôle envers un avatar dont ils partagent le contrôle, ce qui est très prometteur pour de potentielles futures applications de rééducation motrice ou de formation à des gestes techniques. En effet, on peut imaginer un scénario où un patient et médecin pourraient partager un même avatar avec le médecin contrôlant le corps virtuel, encourageant la sensation de mouvement chez le patient.

FIGURE 3. Illustration de l'expérience étudiant l'influence d'un danger virtuel sur le SI [7]. Les utilisateurs étaient amenés à placer une pièce métallique sous une presse qui se déclenchait parfois sur leur bras virtuel.

Menaces dans les EVs

Une autre caractéristique des EVs largement exploitée est leur capacité à transmettre aux utilisateurs un large éventail d'émotions. Pour cette raison, la RV est devenue particulièrement attrayante pour différents domaines de recherche où il est crucial que l'environnement virtuel réussisse à induire des réactions émotionnelles. Cela implique des recherches explorant les réactions émotionnelles des utilisateurs en RV [4] ainsi que des travaux étudiant l'utilisation des menaces virtuelles dans la thérapie d'exposition basée RV pour traiter les phobies [15, 14]. Un autre domaine de recherche qui nous intéresse plus spécifiquement dans ce travail, est l'étude de l'incarnation d'avatars virtuels, où l'introduction d'une menace est fréquemment utilisée pour évaluer le SI des utilisateurs envers leur avatar. Plus précisément, plusieurs études ont montré avec succès que le SI était corrélé à la réaction à une menace virtuelle envers le corps virtuel [17], validant l'hypothèse selon laquelle si les utilisateurs sont bien incarnés dans l'avatar virtuel, ils réagiront physiquement à une menace virtuelle envers leur corps virtuel. Néanmoins, si l'introduction d'une menace virtuelle dans les études d'incarnation virtuelle est largement utilisée, aucune recherche n'a spécifiquement évalué l'impact de la menace virtuelle sur le SI. En d'autres termes, le SI est-il modulé par l'occurrence d'une menace ?

Dans ce contexte, nous avons réalisé une étude [7] qui explore l'impact potentiel de l'introduction d'une menace sur le SI et ne considère donc pas l'introduction d'une menace uniquement comme une mesure, mais comme un facteur susceptible d'affecter le SI. Cette étude explore également l'impact peu connu des répétitions de menaces sur la réaction aux menaces et sur le SI. Nous avons montré que ce type de mesure n'avait pas d'influence sur le sentiment d'incarnation, malgré l'impact potentiel sur l'état émotionnel de l'utilisateur, et qu'elle pouvait donc être utilisée sans risque dans les études sur le SI.

FIGURE 4. Différentes tâches à réaliser dans l'étude explorant l'interrelation des facteurs influençants le sentiment d'incarnation [6].

Influence des facteurs liés aux caractéristiques de l'avatar

Dans une seconde partie, nous avons exploré les interrelations entre les facteurs liés aux avatars qui influencent le SI. Les études explorant l'influence de certains facteurs sur le SI se concentrent généralement sur un facteur à la fois et mesurent son influence sur le SI. Les méthodes utilisées jusqu'alors pour étudier l'influence des caractéristiques des avatars sur la façcon dont ces derniers sont perçus par les utilisateurs ne permettaient pas de savoir si certaines caractéristiques étaient plus importantes que d'autres dans la préférence des utilisateurs. L'évaluation des interrelations est difficile en termes de protocole expérimental en raison des nombreuses combinaisons possibles de facteurs. Pour cette raison, nous étions intéressés dans cette thèse à explorer de nouvelles façons d'évaluer le SI des utilisateurs, et plus spécifiquement d'une manière qui permettrait l'étude des interrelations au sein des facteurs. Nous avons pour cela proposé une méthode innovante [6] pour mesurer l'appréciation des avatars par les utilisateurs qui nous a permis, par exemple, de constater que l'apparence de l'avatar était moins importante que la fidélité et la performance de son contrôle, et que, par conséquent, les designers d'avatars tous domaines confondus gagneraient à allouer leurs moyens dans l'amélioration de cette caractéristique. Nous avons également montré que le type de tâche à effectuer avait un impact sur la préférence des utilisateurs, et devrait donc être pris en compte dans la conception d'avatars en RV.

Influence des facteurs liés à la personnalité des utilisateurs

Si la plupart des études sur l'incarnation d'avatar ont pu montrer des tendances générales concernant la façon dont les facteurs « externes » semblent influencer le SI, elles n'ont pas pris en compte l'aspect « interne » de l'utilisateur (par exemple, la personnalité ou les expériences personnelles). Toutefois, la variabilité entre utilisateurs reste non négligeable. En pratique, on constate que certains croient facilement à l'illusion de l'incarnation virtuelle, alors que d'autres sont au contraire totalement réfractaires. Les premières recherches sur le lien entre les traits de personnalité et

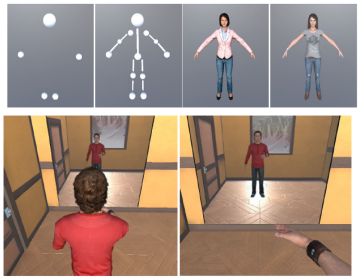


FIGURE 5. Différents niveaux d'apparence (haut) et de point de vue envers l'avatar (bas) dans l'étude explorant l'interrelation des facteurs influençants le sentiment d'incarnation [6].

la perception des expériences de RV se sont concentrées sur le sentiment de présence [16]. Plus récemment, certains travaux ont exploré le lien entre les différences individuelles des utilisateurs et le SI. Par exemple, la conscience du corps [2] et les traits de personnalité [9] ont été étudiés en relation avec le SI. Dans ce dernier cas, les auteurs de [9] ont montré que le sentiment d'agentivité était corrélé au locus de contrôle, un autre trait de personnalité. Cependant, à part les travaux de [9], la majorité des travaux traitant de ces facteurs internes se sont principalement concentrés sur le SI des utilisateurs dans le monde physique. C'est pourquoi nous avons souhaité étudier plus en profondeur l'influence d'un plus large éventail de traits de personnalité et de la conscience du corps sur le SI en RV. Dans un troisième temps, nous avons donc cherché à enrichir les connaissances globales concernant les facteurs influençant le SI en nous concentrant sur les différences individuelles entre les utilisateurs. Nous avons par conséquent exploré l'influence potentielle des traits de personnalité et de la conscience corporelle sur le SI. Cette étude [3], réalisée avec une autre chercheuse (Diane Dewez) a permis d'approfondir les liens entre la personnalité et la perception d'un avatar, ce qui est engageant pour de futures recherches s'intéressant à l'impact des différences individuelles sur la perception d'avatar.

Références

[1] C. Brown, G. Bhutra, M. Suhail, Q. Xu, and E. D. Ragan. Coordinating attention and cooperation in multi-user virtual reality narratives. In *IEEE Virtual Reality*, pages 377–378, 2017.

[2] N. David, F. Fiori, and S. M. Aglioti. Susceptibility to the rubber hand illusion does not tell the whole body-awareness story. *Cognitive, affective and behavioral neuroscience*, 141 :297–306, 2014.

[3] D. Dewez, R. Fribourg, F. Argelaguet, L. Hoyet, D. Mestre, M. Slater, and A. Lécuyer. Influence of personality traits and body awareness on the sense of embodiment in virtual reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 123–134, Oct 2019.

[4] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Mühlberger. The impact of perception and presence on emotional reactions : a review of research in virtual reality. *Frontiers in Psychology*, 6 :26, 2015.

[5] R. Fribourg, F. Argelaguet, L. Hoyet, and A. Lécuyer. Studying the sense of embodiment in vr shared experiences. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 273–280. IEEE, 2018.

[6] R. Fribourg, F. Argelaguet, A. Lécuyer, and L. Hoyet. Avatar and sense of embodiment : Studying the relative preference between appearance, control and point of view. *IEEE Transactions on Visualization and Computer Graphics*, 26(5) :2062–2072, 2020.

[7] R. Fribourg, E. Blanpied, L. Hoyet, A. Lécuyer, and F. A. Sanz. Influence of threat occurrence and repeatability on the sense of embodiment and threat response in vr. In *International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments (ISMAR)*, page 9p, 2020.

[8] R. Fribourg, N. Ogawa, L. Hoyet, F. Argelaguet, T. Narumi, M. Hirose, and A. Lécuyer. Virtual co-embodiment : evaluation of the sense of agency while sharing the control of a virtual body among two individuals. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[9] C. Jeunet, L. Albert, F. Argelaguet, and A. Lécuyer. " Do you feel in control ? " : Towards Novel Approaches to Characterise, Manipulate and Measure the Sense of Agency in Virtual Environments. *IEEE Transactions on Visualization and Computer Graphics*, 24(4) :1486–1495, 2018.

[10] V. Kuszter, G. Brunnett, and D. Pietschmann. Exploring stereoscopic multi-user interaction with individual views. In *Int. Conference on Cyberworlds*, pages 101–106, 2014.

[11] A. Sacau, J. Laarni, N. Ravaja, and T. Hartmann. The impact of personality factors on the experience of spatial presence. *Presence*, 2005.

[12] M. J. Schuemie, P. van der Straaten, M. Krijn, and C. A. van der Mast. Research on presence in virtual reality : A survey. *CyberPsychology & Behavior*, 4(2) :183–201, 2001. PMID : 11710246.

[13] S. Sharma and W. Chen. Multi-user vr classroom with 3d interaction and real-time motion detection. In *2014 International Conference on Computational Science and Computational Intelligence*, volume 2, pages 187–192, 2014.

[14] N. Tardif, C.-t. Therrien, and S. Bouchard. Re-examining psychological mechanisms underlying virtual reality-based exposure for spider phobia. *Cyberpsychology, Behavior, and Social Networking*, 22(1) :39–45, 2019. PMID : 30256675.

[15] J. Wald. Efficacy of virtual reality exposure therapy for driving phobia : A multiple baseline across-subjects design. *Behavior Therapy*, 35(3) :621 – 635, 2004.

[16] H. S. Wallach, M. P. Safir, and R. Samana. Personality variables and presence. *Virtual Reality*, 14(1) :3–13, 2010.

[17] Y. Yuan and A. Steed. Is the rubber hand illusion induced by immersive virtual reality ? In *2010 IEEE Virtual Reality Conference (VR)*, pages 95–102, March 2010.

Reconstruction et correspondance de formes par apprentissage

Thibault Groueix [1]

Thibault Groueix a soutenu sa thèse [2] le 22 octobre 2020 à l'université Paris-Est sous la direction de Mathieu Aubry et Renaud Marlet. Vous trouverez ci-dessous un résumé de cette thèse.

La création artistique en 3D numérique est affaire d'experts car elle est coûteuse en temps et nécessite une grande expérience technique. Afin de rendre cet art de niche accessible à tous, le travail de thèse a visé à développer des algorithmes permettant à un amateur d'effectuer des tâches complexes automatiquement. Par exemple, prendre un objet en photo avec un téléphone et éditer ensuite une estimation de son modèle 3D, ou encore modifier l'apparence d'un objet 3D en transférant la texture d'un autre modèle à l'aide d'un simple clic. Pour ce faire, nous proposons de nouvelles méthodes d'apprentissage profond pour modéliser et analyser les formes 3D en se concentrant sur deux tâches clefs : reconstruire un modèle 3D à partir d'une seule image et mettre des modèles 3D en correspondance.

Une méthode de reconstruction 3D à partir d'une seule image (SVR) est un algorithme qui prend en entrée une image et prédit un modèle virtuel en 3D du monde physique qui a engendré cette image. Ce problème remonte aux premiers jours de la vision par ordinateur, et il est très difficile. En effet, plusieurs configurations de formes, de textures et d'éclairages peuvent expliquer la même image et il y a donc une infinité de solutions au problème. La clef pour attaquer ce problème mal posé est donc de formuler des hypothèses sur ce qu'est un environnement 3D réaliste

1. thibault.groueix.2012@polytechnique.org.

2. http://www.theses.fr/2020PESC1024.

DOI:10.48556/SIF.1024.19.89

et d'utiliser ces hypothèses pour sélectionner une solution à notre problème parmi cette infinité. Dans ce travail, nous apprenons ces hypothèses directement à partir de grandes bases de données, au lieu de les concevoir manuellement. L'apprentissage sur les données permet de formuler des hypothèses très puissantes, qui permettent même de reconstruire les parties invisibles des objets. L'intérêt de développer des méthodes de SVR ne s'arrête pas à la modélisation 3D. En robotique par exemple, il est critique que les agents autonomes disposent d'une représentation 3D de leur environnement à partir de leurs capteurs visuels.

La mise en correspondance de formes vise à établir des correspondances entre des objets 3D. Cette tâche se pose traditionnellement sous forme d'un problème d'optimisation non-convexe, dont la convergence est sujette aux minimum locaux. Au lieu de cela, nous proposons de résoudre un problème d'apprentissage sur de grands jeux de données. La mise en correspondance de formes a de nombreuses applications en modélisation 3D telles que le transfert d'attribut, le gréement automatique pour l'animation ou l'édition de maillage. C'est l'une des briques élémentaires de l'analyse de forme.

Nous présentons à présent les contributions principales de ce travail pour avancer dans la compréhension et la résolution de ces deux tâches.

La première contribution technique est une nouvelle représentation paramétrique des surfaces 3D, que nous modélisons avec des réseaux neuronaux. Le choix de la représentation des données est un aspect critique de tout algorithme de reconstruction 3D. Jusqu'à récemment, la plupart des approches profondes prédisaient des grilles volumétriques de voxel ou des nuages de points, qui sont des représentations discrètes. Au lieu de cela, nous présentons une approche qui prédit une déformation paramétrique de surface. Notre approche, baptisée AtlasNet (cf. figure 1), est la première approche profonde de SVR capable de reconstruire des maillages à partir d'une seule image sans s'appuyer sur un post-traitement, et peut le faire à une résolution arbitraire sans problème de mémoire. Une analyse plus détaillée d'AtlasNet révèle qu'en sus, par rapport aux autres approches par apprentissage, il généralise mieux sa proposition aux catégories pour lesquelles il n'a pas été entraîné.

Notre seconde contribution principale est une nouvelle approche de correspondance de formes, appelée 3D-CODED. Entièrement basée sur des reconstructions par déformation de surface, la clef de voûte de notre raisonnement est de lier correspondances et reconstructions : nous montrons que la qualité des correspondances prédites dépend de la qualité des reconstructions 3D. Nous mettons ainsi en relation deux grands problèmes de 3D. Sur le plan technique, nous introduisons également une optimisation au moment de l'inférence pour affiner les déformations apprises, combinant ainsi apprentissage et optimisation. 3D-CODED (cf. figure 2) se décline en deux versions. Tout d'abord, pour les humains et d'autres catégories qui diffèrent d'une quasi-isométrie, notre approche peut tirer parti d'un modèle de forme. Pour les autres catégories, celles présentant des variations non isométriques, telles

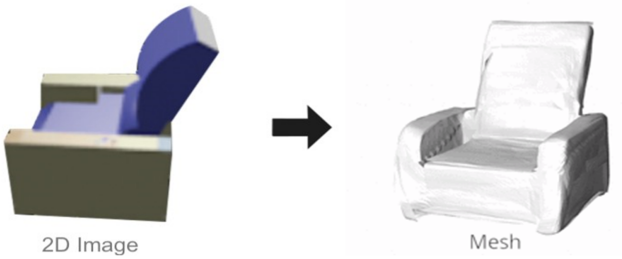


FIGURE 1. Reconstruction 3D à partir d'une seule image : notre approche, AtlasNet, reconstruit un maillage paramétrique de l'objet. Notez que même les parties non visibles dans l'image sont reconstruites.

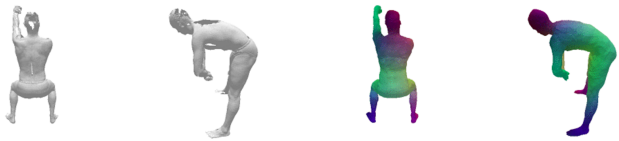


FIGURE 2. Correspondances de formes : notre approche, 3D-CODED, établit automatiquement des correspondances entre deux formes, suggérées par la couleur.

des chaises, nous apprenons à déformer n'importe quelle forme en n'importe quelle autre et introduisons des contraintes de cohérence du cycle pour apprendre des correspondances respectant la sémantique. 3D-CODED fonctionne directement sur les nuages de points ; elle est robuste à de nombreux types de perturbations et a surpassé l'état de l'art de 15 % sur des scans d'humains réels.

Les algorithmes développées dans ce travail ont été présentés en conférences internationales, et ont suscité l'intérêt dans plusieurs groupes de recherches. Le code est libre d'accès sur GitHub [3] et activement utilisé. À ma connaissance, aucun transfert technologique vers un produit commercialisé n'a encore abouti, car de nombreuses questions de recherche restent ouvertes. Toutefois, plusieurs brevets industriels connexes à ces travaux ont été déposés, notamment par l'entreprise Adobe.

3. https://github.com/ThibaultGROUEIX.

Preuves de protocoles cryptographiques : méthodes symboliques et attaquants puissants

Charlie Jacomme [1]

Charlie Jacomme a soutenu sa thèse [2] le 16 octobre 2020 à l'université Paris Saclay, sous la direction de Hubert Comon et Steve Kremer. Vous trouverez ci-dessous un résumé de cette thèse.

Nos ordinateurs, téléphones et autres objets connectés font maintenant partie intégrante de nos vies. Nous les utilisons pour de nombreuses tâches, que ce soit pour communiquer avec nos proches, faire des achats en ligne, prendre des rendez-vous médicaux... Ces systèmes informatiques manipulent alors un grand nombre de nos données personnelles.

Si certaines fuites de données vont clairement à l'encontre de nos intérêts, par exemple lorsqu'un tiers peut obtenir nos informations de carte bleue, les fuites de données telles que notre position ou nos messages représentent des atteintes moins évidentes à la vie privée. On trouve parmi les exemples les plus dramatiques le récent suivi des Ouïghours via leurs téléphones par les autorités chinoises. Il existe aussi des exemples plus subtils : par exemple où une personne peut se voir refuser un prêt car la compagnie d'assurance a pu découvrir qu'elle avait une longue maladie. Il existe ainsi de nombreux exemples où nos données personnelles sont utilisées à

1. charlie.jacomme@cispa.de.
2. https://theses.fr/2020UPASG005.

DOI:10.48556/SIF.1024.19.93

l'encontre de nos intérêts, que ce soit par une personne tierce, un gouvernement [3] ou une entreprise. Une question essentielle se pose alors :

Pouvons-nous avoir des garanties sur l'accès et l'utilisation de nos données ?

La réponse à cette question est aujourd'hui non, et trois aspects doivent être étudiés pour changer cette réponse :

(1) mettre au point des systèmes visant à protéger la vie privée de leurs utilisateurs ;

(2) s'assurer que ces protections sont efficaces ;

(3) et faire en sorte que ces systèmes soient déployés à grande échelle.

Malheureusement, si les entreprises s'intéressent aux questions de sécurité, elles s'intéressent souvent peu aux questions de vie privée, et ce seulement lorsque cela peut compromettre leurs intérêts économiques. Ainsi, il appartient à la recherche publique de se saisir de ces questions, en développant, vérifiant et rendant accessibles des systèmes respectueux de la vie privée, et c'est ce qui m'a intéressé lors de ma thèse. Je me suis intéressé plus particulièrement à l'une des grandes questions scientifiques autour de ce sujet :

Comment fournir des garanties de respect de la vie privée par les systèmes informatiques ?

Contexte scientifique

Les systèmes de communication reposent sur des protocoles de sécurité, i.e. des programmes distribués qui visent à permettre la communication de manière sécurisée entre plusieurs agents. Parmi les protocoles les plus connus et utilisés, on trouve TLS (utilisé pour les connections *https*), SSH ou encore AKA pour les communications téléphoniques.

Ces protocoles reposent souvent sur des primitives cryptographiques telles que le chiffrement (RSA, AES...) ou les fonctions à sens unique (MD5, SHA-3...), qui sont des constructions mathématiques permettant notamment de masquer le contenu d'un message. Par exemple, étant donné une clé secrète sk_A et sa clé publique associée pk_A, une fonction de chiffrement asymétrique `enc` permet à toute personne possédant la clé publique de chiffrer des messages de telle sorte que seule la personne possédant la clé secrète puisse les déchiffrer. Équipé d'une telle primitive, on peut alors imaginer un protocole d'échange de clé entre un initiateur I et un répondeur R qui souhaitent se mettre d'accord sur un secret commun. Pour ce faire, ils peuvent procéder comme illustré dans la figure 1, où chacun tire au hasard une suite de bits dénotée e_X et chiffre cette *bitstring* avec la clé publique de l'autre partie. Les deux parties utilisent alors la valeur du `xor` de ces deux *bitstrings*, $e_I \oplus e_R$, comme nouvelle clé commune. Si l'on considère ce protocole, il est nécessaire de supposer que

3. Même proche de nous, voir par exemple `https://www.lemonde.fr/international/article/2022/01/12/une-application-anti-covid-utilisee-dans-une-enquete-policiere-cible-de-vives-critiques-en-allemagne_6109148_3210.html`.

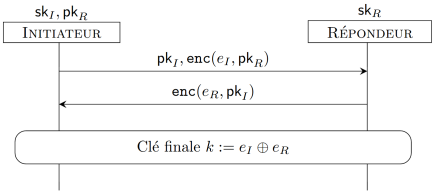


FIGURE 1. Exemple d'échange de clé

le chiffrement fournit certaines garanties pour montrer que la clé finale est effectivement secrète. Ainsi, une première étape pour fournir des garanties sur les systèmes de communication est donc de vérifier que ces protocoles sont sécurisés en supposant que les primitives le sont. Pour cela, il faut modéliser les protocoles, les propriétés de sécurité et les attaquants. C'est là qu'est la difficulté majeure, et l'une des différences avec la vérification classique de systèmes : on doit modéliser de manière réaliste une classe d'attaquants possibles, et montrer qu'aucun attaquant ne peut compromettre le système.

Modèles symbolique et calculatoire Aujourd'hui, deux modèles sont largement utilisés pour cela :

— le modèle symbolique, proposé par [3], où l'attaquant contrôle toutes les communications du réseau, les messages sont modélisés par des termes, et le pouvoir de l'attaquant pour manipuler les messages est capturé par une théorie équationnelle et des règles d'inférence ;

— le modèle calculatoire, initié par [4], où l'attaquant contrôle toutes les communications, peut être n'importe quelle machine de Turing s'exécutant en temps probabiliste polynomial, et les messages sont directement des suites de bits.

Quel que soit le modèle, les preuves peuvent être très complexes et très longues déjà sur des protocoles simples, et réaliser des preuves correctes à la main devient vite difficile, voire impossible. De nombreux outils ont alors été développés pour faire des preuves interactives ou automatiques. Les outils du modèle symbolique (TAMARIN, PROVERIF...) ont le plus haut niveau d'automatisation et permettent de vérifier des protocoles entiers avec des scénarios de compromission complexes. Les outils du modèle calculatoire (EASYCRYPT, CRYPTOVERIF, SQUIRREL...) sont quant à eux moins automatisés. Ils permettent seulement de vérifier des protocoles plus petits, avec des scénarios de compromissions moins complexes mais avec un attaquant plus réaliste. Il existe, de manière générale, un compromis entre le réalisme des modèles et la puissance de calcul de l'attaquant, et les approches symboliques et calculatoires sont complémentaires.

Pour répondre à notre problématique principale, nous pensons donc qu'il est nécessaire d'améliorer, de combiner et d'utiliser plusieurs outils de preuve pour vérifier et améliorer des standards de protocoles déployés à grande échelle.

Contributions de ma thèse

Dans ma thèse, nous avons progressé dans cette direction via plusieurs angles :

— **modularité** : nous avons proposé un cadre général de composition qui permet d'obtenir la preuve d'un protocole à partir de la preuve de ses composants ;

— **preuves assistées par ordinateur** : nous avons travaillé sur l'automatisation d'étapes de preuves bas niveau dans le modèle calculatoire à l'aide d'outils symboliques ou mathématiques ; et nous avons développé un nouveau prouveur interactif, SQUIRREL, permettant de fournir des garanties calculatoires sur la sécurité de protocoles complexes et qui a déjà servi à analyser des protocoles que les outils existants ne pouvaient pas considérer ;

— **mise en pratique** : nous avons effectué une étude de cas pratique, en poussant le modèle symbolique dans ses limites en considérant des attaquants aussi puissants que possibles dans le cadre de l'authentification multi-facteurs ; autour de 5 000 scénarios en 6 minutes ont été analysés via PROVERIF.

Notre résultat de composition à haut niveau

Pour simplifier la complexité des preuves et les rendre plus facilement réutilisables, de nombreux travaux ont essayé de permettre une analyse modulaire des protocoles. L'une des approches principales est l'*Universal Composability* [2], où l'on démontre qu'un composant est sécurisé quel que soit le contexte. L'inconvénient de cette approche est que devoir prouver la sécurité pour tout contexte implique souvent des hypothèses fortes sur les composants, hypothèses souvent non satisfaites en pratique. Le cadre d'application de cette ligne de recherche se retrouve ainsi limité lorsqu'il s'agit d'analyser des protocoles complexes du monde réel tel que SSH.

Une autre approche vise, au lieu de prouver la sécurité indépendamment du contexte, à prouver la sécurité de composants uniquement pour un ensemble restreint de contextes satisfaisant certaines conditions, et ensuite de prouver que le contexte concret satisfait ces hypothèses. Cependant, dans cette ligne de travail, les résultats existants ne s'appliquent qu'à un type de protocole précis, tels que les échanges de clés présentés précédemment, et ne sont pas génériques.

Pour répondre à ces limitations, notre objectif dans ce travail a été de développer un cadre de preuve générique et modulaire où l'on peut prouver qu'un composant est sécurisé dans un ensemble restreint de contextes satisfaisant un certain nombre de conditions, et il suffit de prouver que le contexte du cas d'application satisfait lesdites conditions pour obtenir la sécurité globale.

Nous souhaitions par ailleurs obtenir un résultat de composition qui soit applicable dans le cadre de la logique BC [1], une logique du premier ordre qui permet

de prouver la sécurité des protocoles en obtenant des garanties calculatoires. Si cette logique a rapidement montré son intérêt, et a, par exemple, servi à faire une analyse formelle de AKA à la main, l'une des faiblesses de cette logique était en effet de ne pouvoir considérer qu'un nombre fixe d'agents en parallèle ; faiblesse qu'un résultat de composition peut alors permettre de résoudre en réduisant la sécurité de plusieurs sessions en parallèle à la sécurité d'une unique session. Par ailleurs, cette logique étant la base de SQUIRREL, cela permet alors d'avoir un résultat de composition utilisable dans notre assistant de preuve.

Principale difficulté À un haut niveau d'abstraction, nous considérons des programmes distribués. Si ces programmes ne partagent aucune ressource, il est alors facile de montrer que l'exécution d'un programme n'affecte pas celle des autres, et que l'on peut alors faire des preuves sur chaque programme isolément puis en les combinant ensemble. Par contre, si deux programmes concurrents partagent un même état, cela n'est plus possible. Dans le cadre de la sécurité, si deux protocoles révèlent chacun une moitié distincte d'un secret, chaque programme en isolation ne révèle pas le secret mais le secret sera bien sûr compromis lorsque les deux protocoles s'exécutent en parallèle. Ainsi, la difficulté principale d'un résultat de composition est la capacité à gérer ces ressources que peuvent partager différents programmes.

Notre approche Soit P, Q deux programmes partageant une dépendance commune à une ressource s, par exemple un état ou une clé secrète. On souhaite démontrer que pour tout attaquant A sans accès à s, l'exécution en parallèle des programmes et de l'attaquant vérifie une propriété de sécurité φ qui ne parle que de P, cela dénoté $\forall A.\ (A||P||Q) \models \varphi$. On souhaiterait via un résultat de composition pouvoir faire une preuve qui ne considère pas toutes les exécutions possibles de Q, et se concentre sur P. Notre idée de base est de faire une preuve pour un attaquant A ayant un accès restreint à s lui permettant de simuler le comportement de Q. Si par exemple tous les accès à s dans Q sont de la forme $s := s+6$ ou $s := s+3$, alors il suffit d'augmenter le pouvoir de l'attaquant en lui donnant un moyen d'effectuer à volonté l'opération $s := s+3$ pour qu'il puisse simuler Q. Et une preuve de sécurité contre un tel attaquant est alors une preuve valide pour une exécution dans le contexte de Q.

Nous avons ainsi défini l'idée de simulation par oracle, où l'on donne à l'attaquant accès à un oracle $\mathcal{O}_s$, et on demande à l'attaquant de simuler Q à l'aide de cet oracle. Et ainsi, on peut prouver $\forall A.\ (A||P||Q) \models \varphi$ en démontrant que pour tout attaquant $A^{\mathcal{O}_s}$ ayant accès à l'oracle, $A^{\mathcal{O}_s}||P \models \varphi$. L'idée de la preuve étant que s'il existe un attaquant A et une exécution de P et Q en parallèle violant la propriété, alors il existe également un attaquant $A^{\mathcal{O}_s}$ qui va pouvoir simuler parfaitement cette exécution en interagissant uniquement avec P et donc violer la propriété.

Notre théorème principal établit que pour tout Q tel que la seule ressource commune avec P est s, si $\forall A^{\mathcal{O}_s}.A^{\mathcal{O}_s}||P \models \varphi$ et si Q est simulable grâce à $\mathcal{O}_s$, alors nous

avons que $\forall A.(A||P||Q) \models \varphi$. Nos résultats sont plus généraux, et permettent de couvrir la composition parallèle comme illustré ici, mais aussi la composition séquentielle et surtout la réplication, où l'on considère en parallèle un nombre non borné de copies d'un protocole.

Application à la preuve de sécurité L'approche de la logique BC a été « d'axiomatiser » ce que l'adversaire ne peut pas faire, ce qui a de nombreux avantages et permet de se détacher des détails des modèles d'attaquants et de dériver la propriété de sécurité de la formalisation du protocole et de ces axiomes. Cela permet d'obtenir des résultats dans le modèle calculatoire, dont l'approche est justement de définir ce que l'attaquant ne peut pas faire, e.g., qu'un problème est calculatoirement dur, plutôt que de définir l'attaquant par ce qu'il peut faire comme le fait le modèle symbolique. Le résultat de composition ci-dessus s'adapte alors parfaitement, car il suffit maintenant de considérer des axiomes qui sont corrects pour un adversaire ayant accès à l'oracle $\mathcal{O}_s$. Notre deuxième résultat théorique consiste à montrer concrètement que de tels axiomes peuvent être facilement dérivés des axiomes standards de la cryptographie, par exemple pour les propriétés d'intégrité des signatures digitales.

Enfin, nous avons montré des exemples concrets d'applications pour les protocoles d'échanges de clés. C'est un cas classique de composition, où par exemple après les opérations de figure 1, la nouvelle clé commune peut être utilisée par un autre protocole entre I et R qui est alors composé séquentiellement avec l'échange de clé. Plus spécifiquement nous avons pu, grâce à notre résultat, effectuer la première preuve de sécurité de SSH [5] avec des capacités de forwarding agent pour un nombre arbitraire de sessions.

Références

[1] Gergei Bana and Hubert Comon-Lundh. A Computationally Complete Symbolic Attacker for Equivalence Properties. In Gail-Joon Ahn, Moti Yung, and Ninghui Li, editors, *Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS'14)*, pages 609–620, Scottsdale, Arizona, USA, November 2014. ACM Press.

[2] Ran Canetti. *Universally Composable Security : A New Paradigm for Cryptographic Protocols*. 2000.

[3] D. Dolev and A. Chi-Chih Yao. On the security of public key protocols. In *22nd Annual IEEE Symposium on Foundations of Computer Science, FOCS 1981*, pages 350–357. IEEE Computer Society, 1981.

[4] Shafi Goldwasser and Silvio Micali. Probabilistic encryption. *J. Comput. Syst. Sci.*, 28(2) :270–299, 1984.

[5] Tatu Ylonen and Chris Lonvick. The Secure Shell (SSH) Transport Layer Protocol.

Projet ANR (2015-2018) « Autour du plan 2D »

Julien Castet [1], Florent Cabric [2], Adrien Chaffangeon [3], Dominique Cunin [4], Emmanuel Dubois[2], Elio Keddisseh[2], Yann Laurillau[3], Laurence Nigay[3], Michael Ortega[3], Gary Perelman[2], Carole Plasson[3], Mathieu Raynal[2], Houssem Saidi[2], Marcos Serrano[2]

Introduction

Des situations industrielles à des situations grand public, les contenus numériques sont aujourd'hui au cœur d'une majorité des activités humaines. Les itérations nécessaires à la résolution des problématiques de conception industrielle sont accélérées par la simulation numérique. Les passants sont assistés dans les tâches quotidiennes par la diffusion d'informations géolocalisées (plan de musée, de campus ou de villes). Ces nouveaux usages s'accompagnent d'une mutation de l'activité humaine pour intégrer ces solutions numériques mais également d'une mutation du contenu numérique pour l'intégrer à l'environnement physique des utilisateurs. Ainsi une grande partie des applications actuelles font une place importante aux données 3D et la collaboration autour des contenus numériques (2D et 3D) prend donc une importance grandissante. Les acteurs sont amenés à échanger de l'information, annoter une information existante, co-construire une solution. L'émergence actuelle de ces environnements numériques est donc favorisée d'une part par la forte croissance

1. Immersion, julien.castet@immersion.fr.
2. Université de Toulouse - IRIT, prenom.nom@irit.fr.
3. Université Grenoble Alpes - LIG, prenom.nom@univ-grenoble-alpes.fr.
4. École superieure d'art et design Grenoble-Valence, dominique.cunin@esad-gv.fr.

DOI:10.48556/SIF.1024.19.99

des capacités de calcul et d'autre part par l'amélioration accrue des techniques de visualisation et d'interaction. Pour accompagner ces évolutions, de nouvelles modalités d'interaction en entrée et en sortie participent à une meilleure intégration de ces données dans nos quotidiens personnels ou professionnels.

Le projet « Autour du plan 2D » a visé à fournir des solutions cohérentes et génériques pour l'interaction avec ce nouvel environnement numérique. En effet, bien que les solutions logicielles de partage des données soient en pleine expansion, l'exploitation de ce nouvel écosystème souffre souvent de techniques d'interaction difficiles à utiliser en particulier pour la manipulation 3D grand public et d'un manque notable d'interopérabilité en termes d'interaction humain-machine. Le projet a été motivé par le constat que les différents écrans, surfaces d'interaction, données et utilisateurs sont déjà démultipliés dans plusieurs contextes d'usage aujourd'hui, mais peu de solutions permettent d'en tirer parti facilement et efficacement en favorisant les échanges d'informations ou d'expertises. L'interaction se limite souvent au dispositif local et ne dispose pas de solutions globales et intégrées pour l'accès, la manipulation ou encore l'annotation de ces données par des experts et non experts dans un environnement multi-surfaces (mobile, personnel, partagé). L'espace des possibilités d'interaction est très vaste, incluant l'interaction multi-surface, l'interaction 3D, l'interaction tactile et l'interaction collaborative.

Le projet « Autour du plan 2D » avait donc pour objectif de développer de nouvelles façons de visualiser, interagir et plus globalement collaborer autour de l'information numérique présentée sur une surface horizontale ou verticale. « Autour du plan 2D » a mené cette étude en considérant la situation d'interaction la plus courante dans de nombreux domaines applicatifs : un espace 3D sur une surface horizontale ou verticale dédié au référentiel commun des utilisateurs, le plan 2D, et des multiples espaces interactifs 2D/3D manipulés par les utilisateurs pour explorer et annoter le monde numérique associé au plan 2D. Le projet a abordé ces problématiques en considérant le besoin d'une interaction fluide des utilisateurs avec un environnement défini par un accroissement des données numériques et des dispositifs connectés. « Autour du plan 2D » a traité la problématique de recherche d'interaction multi-dispositifs, multimodale et collaborative, tout en visant l'intégration des technologies nécessaires aux échanges d'information et à la reconnaissance des différents dispositifs dans deux cas d'usage que sont la revue de projet architectural et la consultation publique de données 2D et 3D associées à un territoire (un campus).

Contributions conceptuelles du projet

Afin de bien cerner les problématiques d'interaction liées au projet, une partie de nos travaux a consisté à proposer des supports conceptuels permettant de décrire les différents environnements interactifs étudiés.

En premier lieu, les dispositifs physiques utilisés pour interagir dans des espaces d'interaction riches sont souvent issus de la composition matérielle de dispositifs déjà existants. Dans ce contexte, nous avons élaboré un espace de conception décrivant les différentes formes d'assemblages physiques de dispositifs interactifs, DECO (pour Device Composition) [12]. DECO focalise sur les aspects physiques de la composition de dispositifs et s'articule autour de deux axes : d'une part, l'arrangement physique, qui décrit comment les éléments sont combinés physiquement, et d'autre part, la manipulation physique, qui décrit comment l'utilisateur manipule le dispositif résultant. Afin de valider le pouvoir descriptif de cet espace, nous avons classé les dispositifs de l'état de l'art grâce à DECO. La souris hémisphérique Roly-Poly Mouse (RPM) [10, 11], utilisée dans ce projet est notamment issue d'un processus de conception centrée utilisateur en plusieurs itérations grâce à DECO.

Par ailleurs, nous avons étudié les applications utilisant plusieurs surfaces d'affichage et manipulées au moyen de plusieurs dispositifs. Dans ce contexte, la même action peut donc être exécutée de différentes manières, en utilisant différents dispositifs, selon plusieurs paradigmes d'interaction. Pour décrire ces différentes solutions, en incorporant le support d'affichage, l'information et la technique d'interaction, nous avons introduit le concept de trajectoire d'interaction comme moyen pour décrire l'interaction dans de tels environnements. L'objectif de ces travaux est de fournir un support à l'analyse des activités liées à l'utilisation d'applications multi-dispositifs. Nous avons complété cette notion par l'identification et l'expression de mesures supports à l'analyse de la fluidité de l'interaction et nous avons comparé expérimentalement différentes solutions de conception impliquant plusieurs dispositifs [5].

Enfin, ces applications contiennent généralement des informations multimédia et multidimensionnelles qui sont manipulées par des dispositifs différents selon leur nature. Pour prendre en compte la diversité de ces environnements interactifs multimédia riches, nous avons élaboré un espace de conception basé sur l'identification de différentes couches [3]. Cette approche associe chaque couche de l'espace de conception à une catégorie d'information différente, et correspondant à différents objectifs utilisateurs et espaces d'exploration. À chaque couche sont associés un certain type de données et de représentations, un lien avec les couches précédentes et suivantes, ainsi qu'un ensemble de fonctionnalités d'interaction avec ces données et les dispositifs supports. Nous avons également établi le caractère descriptif, comparatif et génératif de cette approche et l'avons appliqué à la description de différents scénarios d'exploration de données développés dans le contexte de consultation publique de données 2D et 3D associées à un campus.

Techniques d'interaction conçues et évaluées

Dans le projet Autour du Plan 2D, nous avons aussi conçu des techniques d'interaction pour répondre à trois situations : dans un environnement combinant plusieurs écrans, avec une maquette physique, et en utilisant un casque de réalité mixte.

Interaction avec des environnements multi-écrans

Lorsque le volume de données à afficher est très conséquent, il est important d'avoir le plus grand espace d'affichage possible, que celui-ci soit caractérisé par une seule grande surface d'affichage ou sous la forme de plusieurs surfaces combinées. Dans ce contexte, nous nous sommes intéressés à la conception et l'évaluation de nouvelles techniques d'interaction afin de faciliter les tâches élémentaires d'interaction telles que la navigation dans l'ensemble des données, puis leur sélection et manipulation.

En premier lieu, nous nous sommes intéressés à la visualisation et l'exploration de données multidimensionnelles. Les environnements multi-écrans facilitent l'utilisation d'interfaces dites overview + detail, qui affichent une vue détaillée d'une partie d'une grande visualisation. Un environnement multi-écrans permet d'avoir plus d'une vue détaillée en simultané grâce aux différents écrans. Cependant, le nombre de vues détaillées influe grandement sur l'interaction : avoir une seule vue détaillée offre un grand espace d'affichage mais ne permet qu'une exploration séquentielle de la vue d'ensemble ; avoir plusieurs vues détaillées réduit l'espace d'affichage dans chaque vue mais permet une exploration parallèle de la vue d'ensemble. Notre travail a consisté à explorer le bénéfice de diviser la vue détaillée d'une interface overview + detail pour manipuler de larges graphes à travers une étude expérimentale [16]. Pour cela, nous avons conçu une interface overview + multi-détails, nommée Split-focus, permettant d'avoir une vue d'ensemble sur un grand écran et plusieurs vues détaillées (1,2 ou 4) sur une tablette.

Au-delà de la visualisation et de l'exploration des données, il est essentiel de pouvoir pointer facilement l'élément souhaité pour le sélectionner et le manipuler. Le pointage est une tâche élémentaire universelle qui demande à la fois précision et vitesse. Nous nous sommes intéressés au pointage sur un grand écran au moyen d'une montre connectée. Dans ce contexte multi-surface, nous avons étudié si le multiplexage spatial, permettant d'avoir un mode précis et un mode rapide, est viable sur un petit écran de montre. Pour cela l'écran tactile de la montre est découpé en deux zones, avec une zone interne et une zone externe. Nous avons conçu et testé six techniques de pointage. Le multiplexage spatial enrichit le pouvoir d'expression avec l'écran tactile de la montre et autorise une interaction sans regarder l'écran de la montre, mais augmente la charge cognitive de l'utilisateur. De plus, les tests montrent que les gestes effectués sur la montre sont plus lents avec le multiplexage spatial [4].

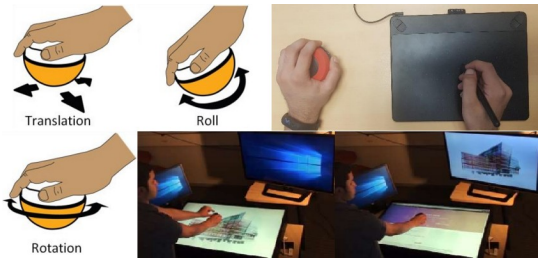


FIGURE 1. A gauche, manipulations possibles avec la RPM : translation, inclinaison et rotation ; à droite, exemples d'utilisation de la RPM sur un écran tactile ou en bimanuel.

Tout au long du projet, nous avons également étudié la souris hémisphérique Roly-Poly Mouse (RPM) qui offre aux utilisateurs cinq degrés de liberté et donne ainsi plus de possibilités d'interaction aux utilisateurs (voir figure 1). Nous avons notamment réalisé Rolling-Menu [6] qui, basé sur l'utilisation de RPM, propose une nouvelle manière d'interagir avec la barre de menu. Rolling Menu contribue à réduire la distance parcourue par le pointeur, due à une nécessaire transition entre le point d'interaction dans l'application et la barre de menus. Il en résulte une meilleure intégration entre la sélection de commande et la manipulation directe du contenu d'une application. Nous avons étudié les inclinaisons de la RPM pour sélectionner un menu ou item dans la barre de menu. De telles manipulations physiques ont l'avantage de nécessiter un temps d'accès équivalent pour chaque menu. Nous avons mené une expérimentation pour évaluer les différentes implémentations de Rolling-Menu et les comparer avec une souris [7]. Les résultats établissent que le mode de validation et la correspondance entre les inclinaisons et les items influencent la performance de Rolling-Menu. Il ressort également que les meilleures techniques de Rolling-Menu sont de 14 % à 40 % plus rapide que la souris pour sélectionner un menu dans une barre de menu contenant de 4 à 10 menus, alors que le taux d'erreur est similaire à celui de la souris.

Nous avons également combiné la Roly-Poly Mouse à d'autres dispositifs d'interaction de manière à augmenter l'espace d'interaction offert aux utilisateurs. Ainsi, nous avons couplé la RPM à un écran tactile ce qui lui permet d'afficher de l'information et de détecter des entrées tactiles. La combinaison de manipulations physiques et de gestes tactiles fait de ce nouveau dispositif, nommé TDome, un dispositif robuste et augmente l'espace de gestes accessibles à l'utilisateur. Ceci lui permet de répondre aux multiples besoins des environnements multi-écrans tout en évitant la multiplication de dispositifs dans l'espace de travail et, notamment, la détection des écrans dans l'espace de travail, la sélection d'écrans, le transfert de données entre écrans et l'interaction avec des écrans distants. L'utilisabilité du dispositif a été évaluée à

travers une étude expérimentale et les résultats montrent que 71 gestes combinés peuvent être confortablement faits avec le dispositif [17].

Nous avons également étudié le couplage de TDome avec l'utilisation d'un stylo numérique utilisé sur une tablette [8]. Le stylet est par exemple utilisé pour annoter des documents. Son utilisation combinée avec le clavier et la souris standard entraîne souvent un coût supplémentaire lors d'une utilisation dans un environnement multi-écran. L'intérêt d'une telle approche est donc de faciliter les tâches inhérentes aux environnements multi-écrans tout en conservant le stylet dans la main dominante, et en facilitant les transitions entre applications des différents écrans au moyen de TDome manipulé par la main non-dominante.

Interaction autour de maquettes physiques

Les métiers du bâtiment ont toujours utilisé des maquettes physiques pour représenter les projets en avance de phase pour avoir ainsi une première représentation des volumes et de la répartition au sol du projet. C'est pourquoi, nous nous sommes également intéressés à l'usage d'une maquette physique et aux interactions tangibles. De même que pour les environnements multi-écrans, nous avons étudié la tâche de pointage avec une maquette physique. Lors d'un usage grand public, les maquettes physiques sont souvent placées derrière une vitrine pour les protéger de la poussière et des dommages. La maquette est alors distante de l'utilisateur. Nous avons conçu et développé une technique utilisant un rayon orthogonal à la surface du verre. Afin d'évaluer l'impact sur la performance de sélection, nous avons mené une expérimentation prenant en compte la taille de la cible, la distance de la cible par rapport à l'utilisateur, la présence de références spatiales et la position de la tête de l'utilisateur par rapport à la vitrine. Les résultats révèlent que l'utilisation du verre comme surface tactile permet de sélectionner facilement des cibles aussi petites que 3 cm jusqu'à 35 cm du verre [2].

D'autre part, les maquettes physiques sont amenées à évoluer au cours de leur utilisation. Afin de pouvoir modifier la maquette physique en cours d'utilisation, nous avons conçu la plateforme EXHI-BIT qui a pour but de prototyper des interfaces déformables mobiles [9]. Elle permet le prototypage d'interfaces déformables 1D, 2D et 3D (voir figure 2). La déformation est manuelle (faite par l'utilisateur) ou automatique (faite par le système grâce à des moteurs). En relation avec le projet, un scénario d'usage dans le cas de l'architecture a été développé et présenté à trois architectures pour commentaires. Le scénario repose sur plusieurs surfaces d'interaction.

Interaction à l'aide d'un casque de réalité mixte

Interactions combinant 2D et 3D Au cours du projet Autour du Plan 2D, nous avons étudié des techniques combinant interaction 2D et 3D sur table à l'aide d'un casque de réalité mixte. Une première étude [13] portant sur le pointage 3D compare les performances de techniques basées sur le casque de réalité mixte et celles basées

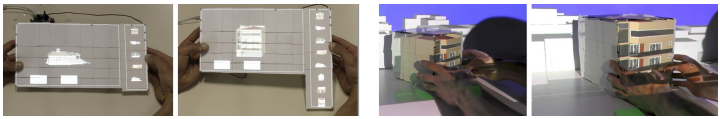


FIGURE 2. EXHI-BIT : prototypage d'interfaces déformables 2D et 3D.

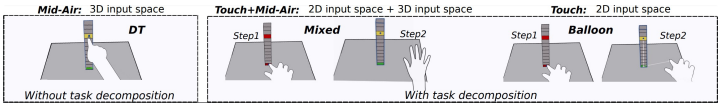


FIGURE 3. Trois techniques conçues et évaluées ; les techniques sont classées selon l'espace d'interaction en entrée (2D ou 3D) et si elles décomposent ou non la tâche.

sur une tablette. Cette dernière offre une surface 2D pour interagir avec un contenu 3D. Nous concluons que les techniques basées sur le casque de réalité mixte, que ce soit par interaction directe ou à distance par technique de lancer de rayon, sont plus appréciées par les participants et requièrent un effort physique moindre.

Nous avons également cherché à comprendre les avantages et inconvénients des trois espaces d'interaction en entrée offerts par la table combinée au casque de réalité mixte [14] : espace 2D (sur la table), espace 3D (au dessus de la table), espace hybride combinant espace 2D et 3D. Trois techniques ont été conçues et développées avec un casque Hololens et une table pour l'interaction tactile (voir figure 3) : une interaction directe dans l'espace 3D, une interaction dans l'espace 2D de la table avec décomposition de la tâche, une interaction avec décomposition de la tâche combinant une interaction dans l'espace 2D puis une interaction 3D dans l'espace au dessus de la table. Appliqué à des tâches de sélection 3D d'objets placés à différentes hauteurs, nous concluons que les interactions avec décomposition de la tâche permettent de gagner en précision grâce au support physique offert par la table. Toutefois, une interaction avec décomposition de la tâche dans l'espace hybride offre un bon compromis entre rapidité et précision.

Enfin, nous avons conçu une nouvelle technique d'interaction nommée RayLens [15] combinant la technique de RayCasting avec un mécanisme de lentille grossissante virtuelle déplaçable dans l'espace 3D. Cette lentille est conçue pour faciliter la sélection d'objets 3D distants. Elle se présente comme un espace d'interaction 2D dans l'espace 3D (voir figure 4), le pointage s'opérant alors sur la surface offerte par la lentille. La comparaison de notre technique RayLens avec la technique classique de Ray-Casting et la technique RaySlider (curseur déplaçable sur le rayon de sélection de la technique RayCasting) montre que la technique RayLens (1) offre

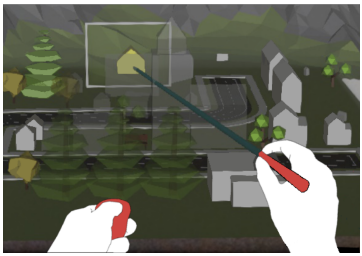


FIGURE 4. Technique RayLens combinant la technique de Ray-Casting avec un mécanisme de lentille grossissante virtuelle déplaçable dans l'espace 3D.

des performances homogènes pour des cibles de petite taille, quelque soit la densité d'objets dans la scène 3D ; (2) induit une charge cognitive et physique réduite.

Exploration collaborative d'une maquette numérique 3D L'usage de casque de réalité mixte pour l'exploration de maquette numérique dans les métiers du bâtiment se développe à grand pas. Une des limitations des écosystèmes mis en place est de manquer de solution efficace pour permettre une exploration immersive et collaborative. Dans la majorité des cas, les solutions proposées se composent de plusieurs casques chaussés par chaque utilisateur, et d'avatar permettant de se retrouver immergés à plusieurs dans la maquette numérique 3D. Pourtant l'observation de ces usages collaboratifs montre que le besoin dépend fortement du profil utilisateur.

Ainsi l'immersion a une vraie valeur ajoutée pour le client mais pas nécessairement pour le commercial. Nous avons imaginé une métaphore d'interaction dans lequel ces deux profils disposent d'outils différents. La spécificité est de munir le commercial d'une tablette et d'un moyen simple pour pouvoir guider le client au cours de sa visite sans chausser un casque. D'autre part, face à l'importance de l'usage de maquettes physiques pour l'exploration de bâtiments, nous avons également développé un prototype d'interaction mixant maquette physique, écran semi-immersif et casque de réalité mixte. Il s'agissait de travailler sur les transitions entre la perspective globale offerte par la maquette et l'immersion locale dans la maquette sur un écran courbe. Le travail de cette transition a été réalisé grâce à un casque de réalité mixte permettant d'afficher des informations sur la maquette physique et de guider l'utilisateur vers une immersion à la première personne.

Enfin, une maquette physique du bâtiment à visiter n'est pas toujours proposée, et les utilisateurs ne disposent que d'une version numérique en 3D. Dans ce contexte, nous avons conçu et réalisé une expérience permettant d'expérimenter une interaction directe avec des hologrammes en proposant un repère 3D dans l'espace qui

permet à l'utilisateur d'être informé par rétroaction que sa main se trouve sur l'hologramme. Ce prototype intègre les dispositifs HoloLens, Ultrahaptics et Leap Motion, ce qui permet de fournir un moyen d'interagir directement avec les hologrammes.

Conclusion

Cet article présente des résultats du projet « Autour du plan 2D » sous la forme d'espaces de conception et de techniques d'interaction. Nous rappelons que ce projet avait pour objectif de développer de nouvelles façons de visualiser, interagir et plus globalement collaborer autour de l'information numérique présentée sur une surface horizontale ou verticale, le plan 2D. Les espaces de conception présentés et la variété des techniques d'interaction décrites souligne l'étendue de l'espace des possibilités d'interaction qu'il convient de continuer à explorer.

Remerciements

« Autour du plan 2D » a été financé par l'Agence nationale pour la recherche (ANR-15-CE23-0001).

Références

[1] Jean-Paul Delahaye. Se libérer du Bitcoin, Introduction aux blockchains et aux cryptomonnaies. Dunod, 2022.

[2] Cabric, F.; Dubois, E.; Irani, P.; Serrano, M. Touchglass : Raycasting from a glass surface to point at physical objects in public exhibits. In IFIP Conference on Human-Computer Interaction, 249–269. Springer, 2019.

[3] Celentano, A.; Dubois, E. A layered structure for a design space dedicated to rich interactive multimedia content. Multimedia Tools and Applications, 76(4), 5191–5220, 2017.

[4] Chaffangeon, A.; Nigay, L.; Laurillau, Y. Pointage sur un grand écran avec une montre connectée multiplexage spatial. Rapport de L3. ENS Rennes, 2017.

[5] Dubois, E.; Celentano, A. Analysing interaction trajectories in multi-device applications. In Proceedings of the 9th Nordic Conference on Human-Computer Interaction, 1–6, 2016.

[6] Dubois, E.; Raynal, M.; Serrano, M. Améliorer l'interaction avec une barre de menus grâce à des gestes d'inclinaison sur une souris multidimensionnelle. In Proceedings of the 29th Conference Francophone sur l'Interaction Homme-Machine, IHM'17, 83–92. ACM, New York, NY, USA, 2017.

[7] Dubois, E.; Serrano, M.; Raynal, M. Rolling-menu : Rapid command selection in toolbars using roll gestures with a multi-dof mouse. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–12, 2018.

[8] Keddisseh, E.; Serrano, M. Dubois, E. Une approche bimanuelle basée sur un stylet pour l'interaction dans des environnements multi-écrans. In Proceedings of the 30th Conference Francophone sur l'Interaction Homme-Machine, IHM'18, 195–201. ACM, New York, NY, USA, 2018.

[9] Ortega, M.; Maisonnasse, J.; Nigay, L. Exhi-bit : a mechanical structure for prototyping expandable handheld interfaces. In Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, 1–11, 2017.

[10] Perelman, G. ; Serrano, M. ; Raynal, M. ; Picard, C. ; Derras, M. ; Dubois, E. Conception d'un dispositif pour interagir avec des données multidimensionnelles : Disco. In Proceedings of the 26th Conference Francophone sur l'Interaction Homme-Machine, IHM'14, 91–100. ACM, New York, NY, USA, 2014.

[11] Perelman, G. ; Serrano, M. ; Raynal, M. ; Picard, C. ; Derras, M. ; Dubois, E. The roly-poly mouse : Designing a rolling input device unifying 2d and 3d interaction. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 327–336, 2015.

[12] Perelman, G. ; Serrano, M. ; Raynal, M. ; Picard, C. ; Derras, M. ; Dubois, E. Deco : A design space for device composition. In Proceedings of the 2016 ACM Conference on Designing Interactive Systems, 435–446, 2016.

[13] Plasson, C. ; Cunin, D. ; Laurillau, Y. ; Nigay, L. Tabletop ar with hmd and tablet : A comparative study for 3d selection. In Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces, 409–414, 2019.

[14] Plasson, C. ; Cunin, D. ; Laurillau, Y. ; Nigay, L. 3d tabletop ar : A comparison of mid-air, touch and touch+mid-air interaction. In Proceedings of the International Conference on Advanced Visual Interfaces, AVI '20, 2020.

[15] Plasson, C. ; Cunin, D. ; Laurillau, Y. ; Nigay, L. A lens-based extension of raycasting for accurate selection in dense 3d environments. In C. Ardito ; R. Lanzilotti ; A. Malizia ; H. Petrie ; A. Piccinno ; G. Desolda ; K. Inkpen, eds., Human-Computer Interaction – INTERACT 2021, 501–524. Springer International Publishing, Cham, 2021.

[16] Saidi, H. ; Serrano, M. ; Dubois, E. Investigating the effects of splitting detailed views in overview+ detail interfaces. In Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services, 180–184, 2016.

[17] Saidi, H. ; Serrano, M. ; Irani, P. ; Dubois, E. Tdome : a touch-enabled 6dof interactive device for multi- display environments. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 5892–5904, 2017.

Projet ANR (2016-2021) « PractiKPharma » : extraction, comparaison et découverte de connaissances en pharmacogénomique

Pierre Monnin [1] et Adrien Coulet [2]

Introduction

Le projet ANR PractiKPharma [3] (2016 – 2021) s'est intéressé au développement d'approches informatiques pour le domaine de la pharmacogénomique (PGx). Ce domaine étudie l'influence des variations génétiques des individus sur leur réponse aux médicaments. En d'autres termes, la PGx s'intéresse aux relations illustrées par la figure 1 qui lient un ensemble de médicaments, un ensemble de facteurs génétiques, et un ensemble de réponses aux médicaments (effets attendus, indésirables, ou absence d'effet). Par exemple, les relations PGx représentées en figure 2 indiquent qu'un individu traité avec de la codéine pourra connaître une absence d'effet, l'effet analgésique attendu, ou un effet indésirable de toxicité en fonction du variant génétique porté par le gène CYP2D6.

L'état de l'art en PGx est conséquent mais très inégalement validé car une grande partie des observations sont peu ou pas reproduites. Par exemple, la figure 3 illustre que 90 % des relations PGx représentées dans PharmGKB [23], base de données

1. Chercheur, Orange, Belfort, pierre.monnin@orange.com.

2. Chercheur, équipe HeKA, Inria Paris, centre de recherche des Cordeliers, Inserm, université de Paris, adrien.coulet@inria.fr.

3. http://practikpharma.loria.fr

DOI:10.48556/SIF.1024.19.109

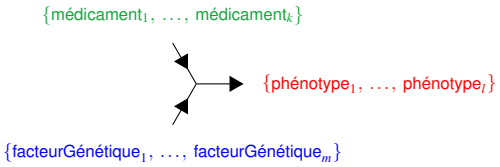


FIGURE 1. Modèle abstrait d'une relation pharmacogénomique.

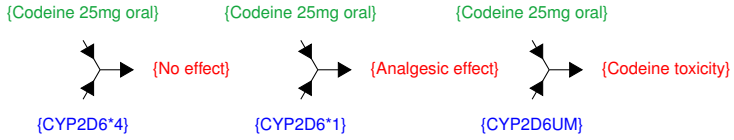


FIGURE 2. Relations pharmacogénomiques décrivant l'influence de trois variants du gène CYP2D6 sur la réponse à un même traitement de codéine. Les patients ayant le variant CYP2D6*4 ne connaîtront pas l'effet analgésique attendu du traitement ; ceux ayant le variant CYP2D6UP connaîtront une toxicité du traitement.

de référence du domaine, ne sont pas directement applicables en pratique clinique par manque de validation. Le but de PractiKPharma a été de fournir des méthodes et outils de gestion de connaissances pour progresser dans la validation de ce qui ne l'est pas, en extrayant et comparant des connaissances issues de sources diverses comme les bases de données spécialisées, la littérature biomédicale et les dossiers patients informatisés (DPI). Dans cet objectif, le projet a suivi quatre axes de travail illustrés en figure 4 et décrits ci-dessous :

(1) l'extraction de connaissances de l'état de l'art à partir de bases de données spécialisées et de la littérature ;

(2) l'extraction de connaissances « observationnelles » à partir des DPI pour identifier les connaissances pouvant être mises en œuvre en médecine personnalisée ;

(3) la comparaison des connaissances extraites en (1) et (2) ;

(4) la valorisation des connaissances extraites en cherchant des relations gènes–médicaments plus probables et des mécanismes moléculaires capables d'expliquer la survenue d'effets indésirables. Ce dernier objectif est motivé par le fait que les mécanismes mis en jeu dans les réponses aux médicaments sont très souvent inconnus. Or, la connaissance des facteurs génétiques associés aux réponses peut servir de point d'entrée à l'identification des mécanismes moléculaires sous-jacents.

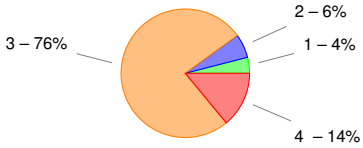


FIGURE 3. Répartition des niveaux de validation des connaissances pharmacogénomiques dans la base de référence PharmGKB (au 05/07/2019). Les niveaux 1 et 2 correspondent à des connaissances implémentées en pratique clinique ou supportées par des études montrant un niveau fort (1) ou modéré (2) d'association. Les niveaux 3 et 4 correspondent quant-à-eux à des connaissances décrites dans des études non-répliquées, dans de multiples études montrant un manque de preuve, ou dans des études non-significatives. 90% (3+4) des connaissances nécessitent plus de validation pour pouvoir être utilisées en clinique.

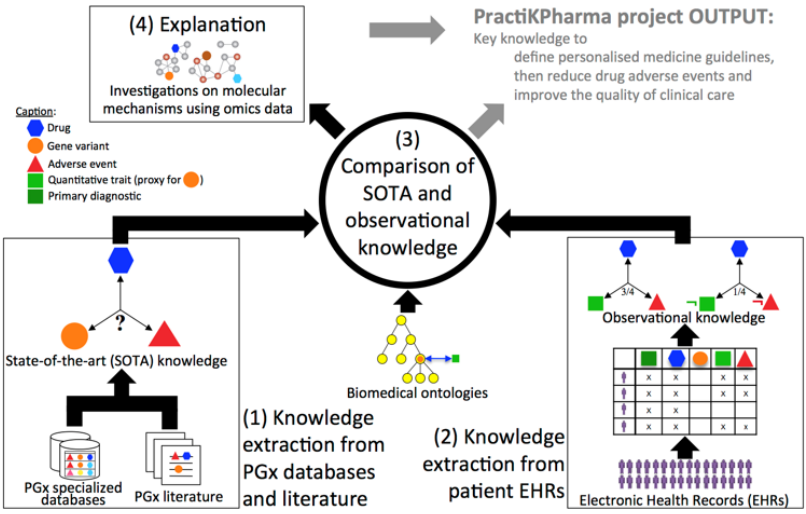


FIGURE 4. Les quatre axes de travail du projet PractiKPharma

PractiKPharma a été mené par :

— des experts en pharmacie et pharmacovigilance du service SSPIM du CHU de Saint-Étienne et du centre régional de pharmacovigilance du CHRU de Nancy ;

— des experts en gestion de dossiers patients informatisés de l'hôpital Georges Pompidou de l'AP-HP ;

— des experts en gestion de connaissances et d'ontologies du laboratoire LIRMM de Montpellier ;

— des experts en gestion et extraction de connaissances du Loria de Nancy.

Principaux résultats

Extraction de connaissances pharmacogénomiques à partir de l'état de l'art

L'objectif de cet axe de travail est de permettre une extraction automatique des connaissances de l'état de l'art en pharmacogénomique (PGx) et de les intégrer dans une plateforme qui permet leur comparaison et leur analyse. Nous avons extrait des connaissances de deux types de sources : des textes, avec les résumés d'articles scientifiques de la base de données PubMed, et des données structurées et parfois semi-structurées de la base de données PharmGKB [23], référence en PGx. Les connaissances extraites (et les entités impliquées comme les gènes et les phénotypes) ont d'abord été normalisées à l'aide de vocabulaires spécialisés du domaine (ici appelés ontologies) avant d'être intégrées dans PGxLOD [4] [16], un graphe de connaissances que nous avons construit en suivant les principes du Web Sémantique [3] et les principes FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) [24].

Il est important de noter qu'au commencement du projet PractiKPharma, il n'existait pas de corpus annoté dédié au domaine de la PGx. En particulier, il n'existait pas de corpus avec des phrases simultanément annotées par les trois types d'entités d'intérêt en PGx : facteur génétique (gène, variant, haplotype, etc.), médicament, et phénotype de réponse au médicament. L'absence d'un tel corpus a guidé nos efforts dans deux directions suivies de front :

(1) l'apprentissage par transfert pour l'extraction de relations PGx, avec l'utilisation de corpus existants mais connexes à notre tâche (où, par exemple, seuls un ou deux des types d'entités d'intérêts sont annotés) ;

(2) la création d'un corpus manuellement annoté, dédié à notre tâche et donc incluant les trois types d'entités.

Considérant l'apprentissage par transfert (1) pour la tâche d'extraction de relations à partir de textes, nous avons montré qu'enrichir un corpus cible (où les relations annotées sont celles que l'on cherche à extraire) de petite taille avec un corpus source de plus grande taille (où les relations annotées sont d'un type distinct) améliore les performances, en particulier lorsque le modèle d'extraction de relations

4. https://pgxlod.loria.fr

peut considérer des informations sur la syntaxe des phrases [14]. Nous avons en particulier exploré l'hypothèse selon laquelle les modèles pourraient gagner en performances en généralisant des connaissances sur la syntaxe de l'expression des relations en anglais, même si celles-ci sont de types différents. Cette hypothèse semble valide dans notre contexte applicatif et mériterait davantage d'expérimentations pour être généralisée.

Considérant la constitution d'un corpus inédit (2), nous avons assemblé et ouvert PGxCorpus [5] [13], un corpus de 945 phrases issues de résumés d'articles scientifiques où les entités d'intérêt en PGx et leurs relations ont été annotées manuellement par 11 annotateurs. Le corpus final contient 2875 relations annotées, chacune ayant été vue par au moins quatre annotateurs différents. Nous avons dans un premier temps montré l'utilité de ce nouveau corpus pour l'extraction automatique de relations [13], puis nous avons montré qu'un modèle de type BERT, pré-entraîné avec des textes biomédicaux multi-domaines et réglé finement avec PGxCorpus, permet de dépasser les meilleures performances de l'état de l'art pour la tâche cible [12]. Nous estimons que PGxCorpus permettra d'une part de progresser dans la gestion des connaissances associées à ce domaine, notamment en permettant d'entraîner ou de régler finement des modèles supervisés. D'autre part, PGxCorpus contient plusieurs particularités linguistiques constituant des challenges en traitement automatique de la langue (TAL) comme des entités discontinues, imbriquées, et des relations ternaires. Nous pensons que PGxCorpus permettra d'évaluer des outils de TAL génériques pour l'extraction de ces entités ou relations particulières.

Concernant l'extraction de connaissances à partir de bases de données expertes, nous avons développé des scripts d'extraction automatique des variants, médicaments, et réponses aux médicaments décrits dans PharmGKB, ainsi que de leurs relations disponibles de façon structurée ou semi-structurée dans la base. Cette extraction ne constitue pas une contribution scientifique en tant que telle, mais le résultat de l'extraction vient compléter notre graphe de connaissances appelé PGxLOD et constitue la version la plus récente de PharmGKB en RDF disponible à la communauté depuis que le projet Bio2RDF n'est plus maintenu [11].

L'ensemble des relations extraites de la littérature ainsi que les relations extraites de PharmGKB sont structurées selon une ontologie minimale appelée PGxO et regroupées au sein d'un graphe de connaissances appelé PGxLOD [16]. En plus des relations PGx, PGxLOD regroupe des connaissances associées aux gènes, médicaments et phénotypes en intégrant notamment le contenu des bases de données ClinVar, DrugBank, SIDER, et CTD. PGxLOD est une ressource qui suit les préceptes du Web Sémantique et respecte les principes FAIR. Notons, par exemple, que chaque connaissance représentée dans ce graphe est associée à une provenance bien définie. PGxLOD est indexé dans Google Dataset Search et LOD Cloud. Ce graphe de

5. https://pgxcorpus.loria.fr.

connaissances constitue la plateforme expérimentale du projet pour la comparaison de connaissances PGx de provenances différentes (Section) et la recherche d'éléments explicatifs et mécanistiques quant à la survenue d'effets indésirables (Section).

Extraction de connaissances pharmacogénomiques à partir des dossiers patients

L'objectif initial de cet axe de travail est d'extraire des connaissances pharmacogénomiques (PGx) à partir de dossiers patients informatisés (DPI). Parmi les tâches prévues, nous souhaitions constituer une cohorte de patients et mesurer la variabilité de leur réponse à un médicament particulier. L'absence de données génétiques dans les DPI était un obstacle identifié dès le départ que nous pensions surmonter en mettant en évidence des *surrogates*, c'est à dire des co-variables mesurées de façon routinières dans les DPI et qui puissent être associées statistiquement à des variants génétiques (non mesurés en routine).

L'accessibilité, le contenu et la qualité des données des DPI ont orienté nos contributions vers une extraction de connaissances à partir des textes cliniques associées aux DPI. Pour cela, nous avons avancé le développement d'outils pour la reconnaissance d'entités nommées dans les textes cliniques et pour la détection de leur contexte : sont-elles niées ? Concernent-elles le patient ou un membre de sa famille ? Le présent ou le passé ? Sont-elles dans une zone dupliquée de compte rendu en compte rendu ? French Annotator, NCBO Annotator+, FrenchFast Context et d'autres approches développées dans le cadre du projet permettent de mieux répondre à ces questions en français [9, 15, 21, 22]. Ces éléments contextuels sont cruciaux afin de limiter le nombre de faux positifs que les outils de reconnaissances d'entités nommées produisent s'ils sont utilisés seuls. En parallèle, nous avons évalué et comparé la facilité de réutilisation des suites logiciels standards pour le traitement automatique du langage (TAL) dans le cas particulier des textes cliniques hospitaliers [10] et nous avons participé au développement d'une librairie appelée PyMedExt dont l'objectif est de faciliter le traitement des textes cliniques [6]. Nous avons développé une plateforme à l'aide de Galaxie, un gestionnaire de *workflows* bioinformatiques, qui facilite la gestion de données génomiques des patients par les biologistes moléculaires de l'HEGP [8]. Nous voyons ces différentes contributions comme des briques logicielles nécessaires pour permettre dans le futur l'extraction de connaissances PGx à partir de DPI.

Dans le cadre d'une collaboration avec Stanford, dont les notes cliniques en anglais sont déjà annotées en tenant compte du contexte, nous avons pu mener deux travaux d'analyse autour de la variabilité individuelle de la réponse aux médicaments. Nous avons utilisé des extensions de l'analyse formelle de concepts (AFC) pour trouver des ensembles de réactions indésirables aux médicaments souvent observées chez les mêmes patients [20]. Et, poussé par l'absence de données génétiques, nous

6. Résultat non encore publié, `https://github.com/equipe22/pymedext_core`.

avons expérimenté l'utilisation des changements de doses de traitement comme marqueur des profils de réponse, en faisant l'hypothèse qu'une réduction de dose est le signe d'une surréaction à un traitement. Dans ce contexte, nous avons montré pour une vingtaine de médicaments connus en PGx que nous pouvions, avec des données récoltées de façon routinière dans les DPI, prédire si un patient bénéficierait d'une réduction de dose pour éviter un effet indésirable et cela avant que le médicament ne soit prescrit [5].

Les résultats obtenus sont conséquents et illustrent l'intérêt d'utiliser les DPI. Néanmoins, ils demeurent en décalage avec l'objectif premier, trop ambitieux, d'extraction de connaissances PGx à partir de ces données complexes. En réponse à ce constat, nos contributions ont visé d'une part à faire face au premier challenge que constitue l'utilisation des textes des DPI ; et d'autre part à prendre du recul, pour nous intéresser à la variabilité de réponse aux médicaments de façon plus générale, et pas seulement celle causée par la génétique.

Comparaison de connaissances

L'objectif de cet axe de travail est de développer des méthodes et outils pour permettre la comparaison des connaissances de provenances diverses. Nous avons expérimenté avec la pharmacogénomique (PGx) mais avons apporté une attention particulière à ce que ces méthodes soient les plus générales possibles.

Notre premier travail d'investigation pour la comparaison de connaissances a été de nous assurer que les vocabulaires et ontologies utilisés dans les différentes sources de connaissances considérées pouvaient être alignés. Nous avons observé que dans la plupart des cas, les alignements multilingues proposés dans [1] sont suffisants pour réconcilier nos données issues de l'état de l'art ("annotables" avec des ontologies anglo-saxonnes) et des DPI ("annotables" avec des ontologies francophones). Ces alignements multilingues mettent en évidence la relative pauvreté lexicale des ressources françaises, en comparaison des ressources anglo-saxonnes. Pour cette raison, nous avons réorienté les efforts de PractiKPharma sur le développement de méthodes d'alignement entre ontologies anglo-saxonnes. Dans cette optique, nous avons étudié l'usage de ressources externes (c'est-à-dire d'autres ontologies) pour l'alignement d'ontologies et proposé une méthode qui dépasse l'état de l'art dans ce domaine [2].

La très grande hétérogénéité des sources de données en termes de vocabulaire, granularité et niveau de discours, fait que la normalisation par ontologies des connaissances extraites précédemment n'est pas suffisante pour leur alignement. Par exemple, la figure 5 illustre l'hétérogénéité des connaissances extraites (différentes langues, différents vocabulaires, arguments inconnus) et des alignements attendus entre elles (connaissances identiques, plus spécifiques, similaires à un certain degré). Le second travail d'investigation s'est pour cette raison intéressé à la comparaison des connaissances regroupées au sein du graphe de connaissances PGxLOD, en tirant au maximum parti des connaissances de domaines associées (*i.e.*, des

ontologies). Dans ce contexte, nous nous sommes intéressés au cas particulier de l'alignement de relations *n*-aires dans un graphe de connaissances, de par l'exemple des relations PGx, qui sont définies par l'ensemble des individus qu'elles relient. L'alignement automatique de telles entités au sein d'un graphe de connaissances n'est pas quelque chose de classique et résolu dans l'état de l'art et nous y avons contribué par deux travaux. Notre premier travail propose une approche symbolique non supervisée, à base de règles formelles qui permettent d'identifier si deux individus du graphe qui ont des provenances distinctes sont équivalents, plus spécifiques, ou similaires, et cela au regard de leurs voisinages directs dans le graphe (*i.e.*, les composants associés par la relation *n*-aire) et des connaissances de domaines [17]. Notre second travail propose une approche supervisée numérique, qui apprend à partir des exemples de pairs d'individus du graphe à distinguer ceux qui sont équivalents, plus généraux, ou similaires [19]. Pour cette seconde approche nous avons appris une représentation de nos relations de façon supervisée avec un réseau convolutif de graphe (GCN pour *Graph Convolutional Network*). Suivant cette approche, nous avons particulièrement étudié la façon avec laquelle le GCN pouvait retrouver des types de similarités distincts à partir du voisinage. Ces deux méthodes, l'une symbolique et l'autre numérique, ont la qualité d'être relativement générales et applicables à d'autres domaines que la pharmacogénomique. Du point de vue du cas d'application de notre projet, nous les avons mis en œuvre pour aligner au sein de PGxLOD les connaissances de provenances diverses. Les nouveaux alignements obtenus sont intéressants, premièrement car ils mettent en lumière des relations qui sont décrites dans plusieurs publications mais sont absentes de PharmGKB (soulignant un manque dans la base) ; deuxièmement des relations présentes dans la base, en manque de références scientifiques pour documenter leur niveau de validation, sont mises en relation avec des publications qui mentionnent ces relations.

Dans cet axe de travail, nous avons proposé et mis en œuvre des méthodes originales de comparaison de connaissances et établi un système fonctionnel qui met en évidence les connaissances équivalentes ou similaires proposées dans différentes sources. Ceci est unique et ouvre des perspectives d'analyse de données plus globales et uniques pour le domaine. Parmi les limites qu'il est important de citer, nous n'avons pas comparé l'état de l'art et des résultats d'analyse de cohortes de DPI, ce qui est une tâche complexe. Elle nécessiterait la mise en place d'une source experte qui associe variant génétique et traits phénotypiques observés dans les DPI. Une telle source n'existe pas et les éléments de connaissances qui pourraient la peupler résultent d'analyses d'envergure appelées *phenome-wide association sudies*.

Explication d'effets secondaires aux médicaments

L'objectif de cet axe de travail est la fouille de graphes de connaissances biomédicales pour découvrir des éléments explicatifs de la survenue d'effets médicamenteux indésirables. En particulier, nous pensons que les graphes de connaissances comme

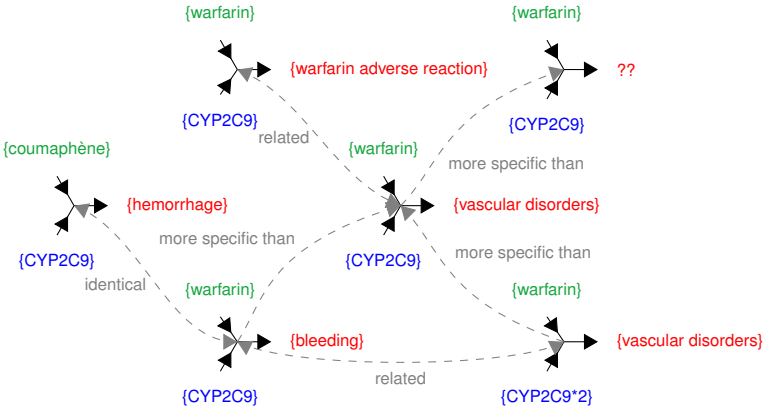


FIGURE 5. Exemples de relations pharmacogénomiques de provenances diverses et des alignements attendus entre elles. Leur hétérogénéité provient des faits suivants : un phénotype est inconnu (??) ; coumaphène est le terme français pour warfarin ; hemorrhage est un synonyme de bleeding ; CYP2C9*2 est un variant génétique plus précis que le gène CYP2C9 lui-même ; bleeding est plus spécifique que vascular disorder ; et vascular disorders est lié à warfarin adverse reaction.

PGxLOD associant des facteurs génétiques aux réponses aux médicaments et incluant de nombreux descripteurs des entités en jeu sont des sources de connaissances sous-exploitées pour la compréhension du mécanisme d'action des médicaments.

Dans un premier temps, PGxLOD a été enrichi avec des sources de données « omiques », telles que KEGG et CTD [16]. La taille de la version résultante de PGxLOD est de 88 millions de triplets [7]. L'exploration d'un graphe d'une telle taille soulève des problèmes de passage à l'échelle, en particulier, lorsque l'on souhaite considérer la sémantique associée au graphe de connaissances. Par exemple, des relations `partOf`, `owl:sameAs`, et `rdfs:subClassOf` sont transitives, et considérer cette propriété lors de la fouille du graphe complexifie la tâche. Pour cette raison, nous avons proposé une approche qui contrôle la quantité de chemins et de patrons de chemin à considérer lorsque l'on explore un graphe de connaissances dans une tâche de fouille [18]. Nous proposons de contrôler la quantité de chemin et patrons de chemin à la fois par des contraintes ad-hoc définies par l'analyste, mais également par des propriétés de monotonicité définies par les données. Nous avons utilisé cette

7. Les triplets ont la forme $\langle$sujet, prédicat, objet$\rangle$ et constituent les éléments de base des graphes de connaissances du Web Sémantique.

approche de fouille pour extraire de PGxLOD des chemins et patrons de chemins associés aux médicaments, et avons montré que ces chemins et patrons de chemins (1) sont de bons prédicteurs pour distinguer les médicaments qui causent un type d'effet indésirable des autres médicaments ; (2) constituent des éléments d'interprétation de la mécanique moléculaire sous-jacente à la survenue d'effets indésirables [4]. Pour cela, nous avons mis en œuvre des méthodes d'apprentissage supervisé symboliques simples mais nativement explicables que sont les arbres de décision et les règles de décision. Parmi les chemins et patrons de chemins qui sont des bons prédicteurs pour la classe de médicaments causant l'effet indésirable, nous sélectionnons ceux qui contiennent une entité de notre graphe potentiellement interprétable par une expert en pharmacologie (par exemple un terme *Gene Ontology*, le nom d'un réseau métabolique, ou celui d'un gène). Les prédicteurs sélectionnés ont ensuite été soumis à des experts en pharmacologie pour évaluer leur potentiel « explicatif », c'est-à-dire leur potentiel pour constituer une explication à la survenue de l'effet indésirable.

Conclusion et perspectives

Le projet PractiKPharma nous a permis de faire progresser l'état de l'art dans le domaine de la gestion des connaissances en pharmacogénomique et de produire des logiciels et ressources offrant des perspectives de nouveaux travaux et collaborations. Nous distinguons trois grands groupes de contributions : des contributions informatiques, d'ordre méthodologique, notamment autour de la comparaison de connaissances de provenances diverses ; des contributions applicatives sur l'extraction de connaissances à partir de textes cliniques avec le développement et partage d'outils fonctionnels ; des ressources ouvertes de référence pour la gestion des connaissances dans le domaine de la pharmacogénomique : un corpus de textes annotés (PGxCorpus) qui permet le développement d'outils d'extraction de connaissances plus performants que l'existant ; et un graphe de connaissances (PGxLOD) qui permet de comparer les connaissances du domaine.

Les ressources constituées au cours du projet ouvrent des perspectives d'un point de vue applicatif, mais également d'un point de vue informatique. En effet, PGxCorpus et PGxLOD offrent des terrains d'expérimentation respectivement pour le traitement naturel de langage (notamment autour des tâches de reconnaissance d'entités imbriquées ou discontinues et d'extraction de relations *n*-aires) et pour la comparaison de connaissances (notamment autour de l'alignement de relations *n*-aires, de l'inférence de liens). D'un point de vue applicatif, ces ressources offrent la possibilité de comparer les connaissances de l'état de l'art et d'identifier les éléments qui nécessitent plus de validation ou, au contraire, d'identifier des faisceaux concordants dans différente sources à propos d'un élément de connaissances, participant ainsi à sa confirmation. En perspective, citons notre volonté de connecter les connaissances de l'état de l'art représentées dans PGxLOD avec des données d'ordre observationnel

concernant des patients suivis à l'hôpital. Ainsi, un ou plusieurs patients ayant vécu une réponse médicamenteuse indésirable viendraient alors instancier une connaissances de l'état de l'art, permettant de valider (ou de modérer) les connaissances de l'état de l'art.

Le récent article de commentaires de Joshua Denny (*Chief Executive Officer of the National Institutes of Health's All of Us Research Program*) du 18 mars 2021 positionne la pharmacogénomique, l'utilisation des DPI, et l'intelligence artificielle comme des éléments clés dans la réalisation de la médecine de précision dans les années à venir [7]. L'expertise acquise dans ces domaines par les partenaires du projet nous positionne de façon favorable pour participer activement à cette réalisation via, notamment, notre participation à la création de l'équipe-projet HeKA (Inria, Inserm, université Paris) [6].

Remerciements

Ces travaux ont été soutenus par l'Agence nationale de la recherche dans le cadre du projet PractiKPharma (ANR-15-CE23-0028), par Inria dans le cadre de l'équipe-associée Inria-Stanford *Snowball*, et par l'IDEX « Lorraine université d'excellence » (15-IDEX-0004).

Références

[1] Amina Annane et al. Réconciliation d'alignements multilingues dans BioPortal. In *IC : Ingénierie des Connaissances*, June 2016.

[2] Amina Annane et al. Building an effective and efficient background knowledge resource to enhance ontology matching. *Journal of Web Semantics*, 51 :51–68, 2018.

[3] Tim Berners-Lee et al. The semantic web. *Scientific american*, 284(5) :28–37, 2001.

[4] Emmanuel Bresso et al. Investigating ADR mechanisms with explainable AI : a feasibility study with knowledge graph mining. *BMC Medical Informatics Decision Making*, 21(1) :171, 2021.

[5] Adrien Coulet et al. Predicting the need for a reduced drug dose, at first prescription. *Scientific Reports*, 8(1), October 2018.

[6] Adrien Coulet et al. L'équipe-projet HeKA. *Bulletin de l'Association Française pour l'Intelligence Artificielle*, pages 29–32, 4 2021.

[7] Joshua C. Denny and Francis S. Collins. Precision medicine in 2030 – seven ways to transform healthcare. *Cell*, 184(6) :1415–1419, 2021.

[8] William Digan et al. An architecture for genomics analysis in a clinical setting using Galaxy and Docker. *GigaScience*, 6(11), November 2017.

[9] William Digan et al. Evaluating the impact of text duplications on a corpus of more than 600, 000 clinical narratives in a french hospital. In *MEDINFO 2019 : Health and Wellbeing e-Networks for All - Proceedings of the 17th World Congress on Medical and Health Informatics*, volume 264 of *Studies in Health Technology and Informatics*, pages 103–107. IOS Press, 2019.

[10] William Digan et al. Can reproducibility be improved in clinical natural language processing ? A study of 7 clinical NLP suites. *Journal of the American Medical Informatics Association*, 28(3) :504–515, 2021.

[11] Michel Dumontier et al. Bio2RDF Release 3 : A larger, more connected network of Linked Data for the Life Sciences. In *Posters & Demonstrations Track, 13th International Semantic Web Conference, ISWC*, volume 1272 of *CEUR Workshop Proceedings*, pages 401–404. CEUR-WS.org, 2014.

[12] Walid Hafiane et al. Expérimentations autour des architectures d'apprentissage par transfert pour l'extraction de relations biomédicales. In *21ème édition de la conférence "Extraction et Gestion des Connaissances", EGC*, January 2021.

[13] Joël Legrand et al. PGxCorpus, a manually annotated corpus for pharmacogenomics. *Scientific Data*, 7(1) :3, 2020.

[14] Joël Legrand et al. Syntax-based transfer learning for the task of biomedical relation extraction. *Journal of Biomedical Semantics*, 12(1) :16, 2021.

[15] Mehdi Mirzapour et al. French fastcontext : A publicly accessible system for detecting negation, temporality and experiencer in french clinical notes. *Journal of Biomedical Semantics*, 117 :103733, 2021.

[16] Pierre Monnin et al. PGxO and PGxLOD : a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. *BMC Bioinformatics*, 20-S(4) :139 :1–139 :16, 2019.

[17] Pierre Monnin et al. Knowledge-based matching of n-ary tuples. In *Ontologies and Concepts in Mind and Machine - 25th International Conference on Conceptual Structures, ICCS*, volume 12277 of *Lecture Notes in Computer Science*, pages 48–56. Springer, 2020.

[18] Pierre Monnin et al. Tackling scalability issues in mining path patterns from knowledge graphs : a preliminary study. In *1st International Conference "Algebras, graphs and ordered sets", ALGOS*, volume 2925 of *CEUR Workshop Proceedings*, pages 123–137. CEUR-WS.org, 2020.

[19] Pierre Monnin et al. Discovering alignment relations with graph convolutional networks : a biomedical case study. *Semantic Web*, pages 1–20, 2021.

[20] Gabin Personeni et al. Discovering associations between adverse drug events using pattern structures and ontologies. *Journal of Biomedical Semantics*, 8(1) :29 :1–29 :13, 2017.

[21] Andon Tchechmedjiev et al. Enhanced functionalities for annotating and indexing clinical text with the NCBO annotator+. *Bioinformatics*, 34(11) :1962–1965, 2018.

[22] Andon Tchechmedjiev et al. SIFR annotator : ontology-based semantic annotation of french biomedical text and clinical notes. *BMC Bioinformatics*, 19(1) :405 :1–405 :26, 2018.

[23] Michelle Whirl-Carrillo et al. Pharmacogenomics knowledge for personalized medicine. *Clinical pharmacology and therapeutics*, 92(4) :414, 2012.

[24] Mark D. Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1) :160018, 2016.

Comment être juste avec des anonymes ?

La mise au point de protocoles anonymes pour les blockchains suggère un intrigant problème de récréation mathématique

Jean-Paul Delahaye [1]

Un débat fait rage depuis une dizaine d'années concernant les méthodes de sécurisation des cryptomonnaies anonymes. Il concerne le Bitcoin et ses concurrents principaux Ethereum, Cardano, Solana et a pris une importance de plus en plus grande. De la conclusion qu'on tire de ce débat dépend sans doute l'avenir des cryptomonnaies et plus généralement des crypto-actifs dont l'impact écologique s'accroît gravement aujourd'hui.

Le problème posé peut s'expliquer en considérant un jeu assez simple qui dégage le problème de son contexte un peu confus. Ce jeu permet à chacun de comparer les deux solutions principales opposées et peut-être d'en imaginer de nouvelles. Voici ce « jeu du mécène et des enveloppes anonymes ».

Distribuer au mieux de l'argent

Imaginez qu'un riche mécène vous confie une certaine somme S, par exemple un million d'euros, en vous demandant de la distribuer. Vous êtes chargé d'organiser une sorte de jeu pour répartir le plus équitablement possible la somme S entre les gens qui acceptent de recevoir une part de cet argent. Vous leur demandez de manifester leur souhait en vous faisant parvenir dans un lieu fixé une enveloppe avec

1. Professeur émérite, université de Lille, campus Cité scientifique, CRIStAL UMR CNRS, 9189 Centre de recherche en informatique signal et automatique de Lille, bâtiment ESPRIT, 59655, Villeneuve d'Ascq Cedex France. E-mail : jean-paul.delahaye@univ-lille.fr.

DOI:10.48556/SIF.1024.19.121

une adresse dans laquelle vous mettrez ce que vous leur attribuez, avant de renvoyer l'enveloppe à l'adresse inscrite. Si vous collectez N enveloppes, la méthode qui vient immédiatement à l'esprit consiste à mettre $\frac{S}{N}$ euros dans chacune des enveloppes et à les renvoyer.

La « méthode S/N » n'est pas parfaite car les personnes ayant plusieurs maisons pourraient mettre plusieurs enveloppes à leur nom avec des adresses différentes ce qui vous empêcherait de savoir que l'argent ira à la même personne. Il y a pire, le mécène vous a demandé de concevoir une méthode qui n'oblige pas ceux qui veulent recevoir une part de S à délivrer leur adresse véritable. Vous acceptez donc que des joueurs se regroupent sur une même adresse et vous soumettent des enveloppes avec des adresses anonymes du type « Joueur 237, 11 rue Dupont à Paris ».

Vous comprenez bien que si vous utilisez la méthode S/N, chaque joueur va être tenté d'envoyer une multitude d'enveloppes dont les contenus lui seront destinés. Il utilisera plusieurs adresses, ou une seule adresse A en prenant le rôle du joueur 1, du joueur 2, etc. de l'adresse A. Si vous utilisez la méthode S/N, vous allez donc recevoir des milliers d'enveloppes, peut-être des millions et les plus malins seront ceux qui auront envoyé le plus grand nombre d'enveloppes pour eux-mêmes et s'approprieront une grande part de la somme S. Les plus gros tricheurs seront favorisés aux dépens de ceux, honnêtes, qui n'auront envoyé qu'une enveloppe. Comment éviter cela ?

Réfléchissez bien, ce n'est pas facile. Il n'existe aucune façon parfaite de résoudre le problème, mais il existe quand même plusieurs façons d'organiser la distribution rendant totalement inopérante la tricherie de l'envoi de plusieurs enveloppes et assurant une certaine équité entre les joueurs. Nous expliquerons plus loin que cette petite énigme mathématique est la transposition du problème du consensus dans les blockchains anonymes de cryptomonnaies.

Solution 1. La méthode de l'enjeu

Vous demandez à chaque joueur de joindre à son enveloppe une certaine somme d'argent. Vous distribuez alors la somme S dans les enveloppes en proportion des sommes qui ont été jointes et en laissant la somme envoyée qui est donc retournée à l'envoyeur. En clair, si l'enveloppe E_i contient s_i, vous laissez s_i dans l'enveloppe, vous y ajoutez $S \times \frac{s_i}{s_1+\ldots+s_n}$, puis vous renvoyez l'enveloppe.

Il est immédiat de vérifier que cela constitue bien une distribution totale de S. Ceux qui ont manifesté un soutient plus fort au jeu et donc une bonne confiance en son organisateur en mettant des sommes importantes dans leur enveloppe sont favorisés proportionnellement au risque pris. Ce n'est pas absurde car justement vous voulez favoriser ceux qui soutiennent le jeu et ont confiance en vous. Le point important est que cette méthode rend inopérant l'envoi de plusieurs enveloppes par une même personne car si, par exemple, quelqu'un envoie deux enveloppes avec 10 euros dans chacune, le gain qu'il en tirera sera le même qu'avec une enveloppe contenant

20 euros. Le gain d'un joueur est proportionnel à ce qu'il est prêt à engager — la somme des s_i de ses enveloppes — et ne dépend pas du nombre d'enveloppes qu'il envoie. Utiliser plusieurs enveloppes plutôt qu'une ne fait rien gagner. Nous nommerons cette organisation de la distribution la « méthode de l'enjeu » car la somme mise par un joueur dans une enveloppe est comme un enjeu qu'il risque puisque l'organisateur pourrait garder l'argent.

Solutions 2. La méthode du travail

Une seconde idée réussit aussi à éviter les tricheurs qui envoient plusieurs enveloppes. L'organisateur de la distribution demande à chaque joueur de réaliser un ruban de dentelle brodée d'un type bien précisé. Le joueur en fabrique la longueur qu'il souhaite, ce qui lui prend du temps, il met le bout de ruban dans l'enveloppe qu'il envoie à l'organisateur du jeu.

Certains joueurs enverront 10 cm de ruban, d'autres joueurs un mètre, etc. La distribution de la somme se fera alors en proportion de la longueur des rubans de chaque enveloppe et les rubans seront détruits. En clair, si l'enveloppe E_i contient un ruban de longueur L_i, après l'avoir vidée de son ruban, l'organisateur y placera la somme $S\frac{L_i}{L_1+\ldots+L_n}$ et la renverra. Comme précédemment, la méthode rend inopérante l'envoi de plusieurs enveloppes par un même joueur car, en envoyant deux enveloppes avec 10 cm de ruban, il gagnera autant qu'en envoyant une seule enveloppe avec 20 cm de ruban. Il y a un certain travail à faire pour réaliser le ruban en dentelle, c'est en quelque sorte une mesure de confiance que le joueur accorde à l'organisateur de la distribution et un engagement de sa part. C'est ce travail qui détermine son gain. Plus un joueur travaille à fabriquer un long ruban, plus il reçoit une part importante de S. On nommera cette méthode « la méthode du travail ».

On peut imaginer d'autres méthodes mais ce n'est pas facile et aucune n'est parfaite. Dans un premier temps, comparons la « méthode de l'enjeu » et la « méthode du travail ».

Comparaison

Posons-nous deux questions. Est-ce que les deux méthodes sont équivalentes l'une à l'autre, ou l'une est-elle meilleure que l'autre ? Que se passe-t-il si la distribution de la somme S est répétée toutes les semaines pendant une longue période ?

La méthode de l'enjeu favorise ceux qui sont assez riches pour joindre à leur enveloppe une somme importante. C'est assez ennuyeux, mais la méthode du travail souffre du même inconvénient, car si vous être riches vous paierez d'autres personnes pour faire de grandes longueurs de ruban qui vous permettront de gagner plus. Sur ce plan, les deux méthodes se valent, elles favorisent ceux qui ont les moyens de s'engager et prennent le risque de le faire.

1. Le jeu du mécène et des enveloppes

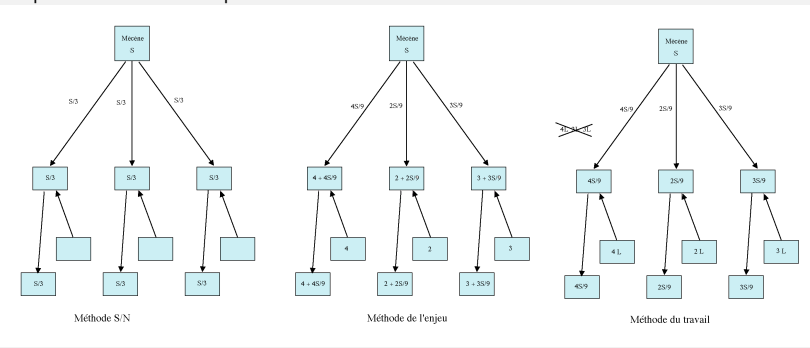
Un mécène veut distribuer une somme S le plus équitablement possible. Il reçoit des enveloppes, il veut répartir S dans les enveloppes.

— Méthode S/N : S'il reçoit N enveloppes, il dépose dans chacune S/N et les renvoie. Ceux qui ont envoyé plusieurs enveloppes sont favorisés.

— Méthode de l'enjeu : Il exige que chacun risque un enjeu en mettant une somme dans son enveloppe. Il renvoie les enveloppes en y laissant l'enjeu, et en y ajoutant une partie de S proportionnelle à l'enjeu trouvée dans l'enveloppe.

— Méthode du travail : Il demande qu'on joigne à l'enveloppe un ruban difficile à fabriquer. Il détruit le ruban, et place dans l'enveloppe une partie de S proportionnelle à la longueur du ruban trouvée dans l'enveloppe.

La question est de savoir qu'elle est la meilleure méthode ?



Quand la distribution des S euros est opérée chaque semaine par la méthode de l'enjeu, il se produit quelques petits changements au cours du temps. Une fois persuadé que le jeu est sérieux, un nombre plus grand de joueurs y participent, certains joueurs empruntent de l'argent pour pouvoir s'engager plus et gagner plus. L'évolution principale est sans doute une augmentation globale des sommes engagées de semaine en semaine.

Si on suppose que la distribution est opérée par la méthode du travail, il se produit quelque chose de très différent, mais assez simple à analyser car il s'agit d'un problème élémentaire d'économie. Certains joueurs développent des moyens pour produire le plus de rubans de dentelle possible. Ils finissent par concevoir des machines spécialisées pour faire efficacement les rubans demandés et réussissent à diminuer les coûts de fabrication. À la longue, certains d'entre eux disposent de multiples machines performantes et fabriquent des longueurs énormes de rubans pourtant destinés à être détruits. Ils deviennent ce que nous appellerons des « industriels des rubans ». Ils sont en compétition les uns avec les autres et les joueurs qui fabriquent les rubans à la main voient leur part de l'attribution diminuer de semaine en semaine et même

devenir négligeable, car les industriels des rubans s'emparent de la presque totalité de *S* à chaque distribution.

La concurrence entre les industriels des rubans engendre un problème de rentabilité. Un industriel ne peut gagner une somme *G* lors de la distribution hebdomadaire de la somme *S*, que s'il dépense pour la fabrication des rubans une somme qui se rapproche de *G*. En effet, si la fabrication pour gagner *G* coûte sensiblement moins que *G*, des concurrents attirés par le bon rendement de ce type d'investissement développent leurs outils et entrent sur le marché ce qui fait baisser le rendement. À la longue, le rendement de la fabrication des rubans diminue et, par exemple, conduit les industriels du ruban à dépenser 80 % (peut-être plus) de la somme *G* qu'ils gagnent chaque semaine en coût de production des rubans. Globalement, à cause de la compétition qui règne entre eux, les industriels des rubans qui se sont emparés de la totalité de la somme *S* distribuée chaque semaine dépensent par exemple 80 % de *S* chaque semaine pour produire les rubans qui leur rapportent donc collectivement 20 % de *S*. Le rendement sous cette hypothèse est de 25 % (20 % comparé à 80 %) ce qui est suffisant pour que l'activité se poursuive. L'équilibre évoluera peut-être et la concurrence suscitera des progrès dans la fabrication des rubans, mais le rendement des investissements des industriels des rubans sera toujours assez réduit, par exemple 25 % ou moins si la concurrence devient plus féroce.

La situation, en fait, est devenue absurde : au lieu de distribuer chaque semaine la somme *S*, si on prend en compte les dépenses pour produire les rubans, chaque

semaine c'est seulement 20 % de S qui est réellement distribuée par le mécène. Non seulement, il y a concentration de la distribution sur les industriels du ruban, mais en plus 80 % de la somme S est en quelque sorte dépensée en pure perte pour fabriquer des rubans aussitôt détruits.

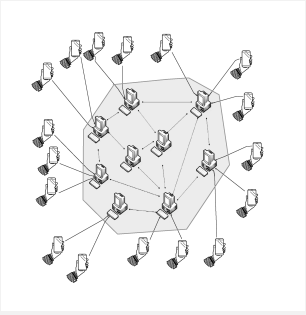
Comparé à la méthode de l'enjeu, la méthode du travail semble catastrophique et le mécène qui donne la somme S chaque semaine ne sera certainement pas satisfait qu'on distribue S par la méthode du travail. Les deux méthodes évitent la tricherie des enveloppes multiples, aucune des deux méthodes n'est parfaite, mais la méthode de l'enjeu est clairement meilleure.

Une variante

On peut aussi imaginer que les sommes s_i ne sont pas rendues mais confisquées. Un problème assez délicat en résulte. Chacun essaiera d'engager le plus d'argent possible — donc en mettant une somme s_i la plus élevée possible dans son enveloppe.

2. Le fonctionnement des crypto anonymes

Une cryptomonnaie s'appuie sur un réseau de validateurs indépendants qui s'occupe chacun de garder le registre des comptes (dénommé « fichier blockchain ») et de le mettre à jour en accord avec les autres validateurs. Pour que la monnaie puisse circuler et qu'elle soit bien sécurisée, il faut qu'il y ait assez de validateurs. On paye donc les validateurs. Ce paiement s'opère souvent en créant des unités de la cryptomonnaie, et en demandant aux utilisateurs du réseau (ceux qui détiennent des comptes) de payer des commissions quand ils font des opérations sur leurs comptes. Ce sont ces nouvelles unités et ces commissions qui constituent la somme S que le réseau distribue aux validateurs. La distribution de S se fait périodiquement, par exemple toutes les 10 minutes pour le réseau Bitcoin. Lorsque les validateurs sont bien identifiés, la distribution de S est facile. Par exemple par la méthode S/N, chacun des N validateurs reçoit $\frac{S}{N}$. Lorsque les validateurs sont anonymes une tricherie est à craindre : que chaque validateur apparaisse plusieurs fois, voire des centaines de fois car l'anonymat rend impossible de savoir si des validateurs apparemment différents le sont réellement. C'est ce qu'on dénomme une attaque Sybil. Pour éviter l'injustice que créent de telles attaques, il faut trouver autre chose que la méthode S/N.



Au centre : le réseau des validateurs, à l'extérieur : les utilisateurs.

Une idée est souvent utilisée. Demander à chaque validateur V_i d'engager une somme s_i en la mettant sous séquestre. La distribution se fait en proportion des s_i. La méthode se dénomme « la preuve d'enjeu ». Satoshi Nakamoto (c'est un pseudonyme), l'inventeur des cryptomonnaies, a proposé la méthode de la « preuve de travail », qui consiste à demander à chaque validateur V_i de réaliser un travail de calcul dont il décide de la puissance c_i. La distribution se fait en proportion des c_i. La preuve de travail et la preuve d'enjeu résolvent le problème des attaques Sybil, mais on a découvert qu'elles produisent des effets économiques très différents.

Mais si le total de l'argent engagé par les joueurs est supérieur à S, alors la somme redistribuée sera inférieure à ce que les joueurs auront engagé et sacrifié ; ils seront donc globalement perdants, et puisque l'argent redistribué à chacun est proportionnel à son enjeu, chaque joueur sera individuellement perdant. Le jeu deviendrait un peu fou car chacun serait tenté d'engager le plus possible en espérant que les autres engagent le moins possible. Une sorte de piège serait tendu aux joueurs qui globalement risqueraient de perdre de l'argent au profit de l'organisateur. Cette méthode pourrait être intéressante pour l'organisateur mais pour notre problème où le mécène veut vraiment donner S, cette option amusante n'a pas d'intérêt.

La variante dans laquelle l'argent des enjeux est donné à une organisation caritative est sympathique. Elle revient à forcer les joueurs à abandonner une part de leurs gains en la donnant à l'organisation caritative, ce qui n'est pas le but du mécène.

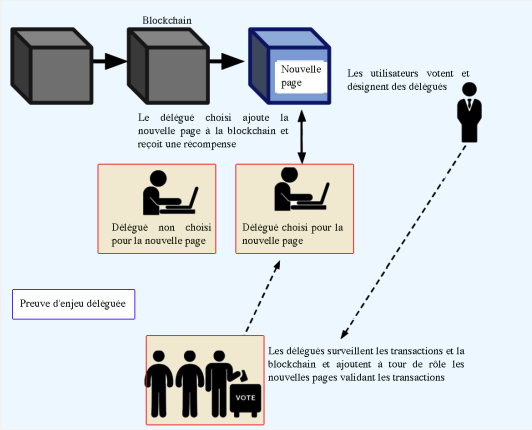
Rien ne semble sensiblement mieux que les deux premières méthodes décrites, mais si des lecteurs en imaginent d'autres plus satisfaisantes, ils doivent me le signaler.

Variantes probabilistes

La méthode de l'enjeu et la méthode du travail ont chacune une variante probabiliste essentiellement équivalente. Il faut les expliquer car elles vont nous approcher

3. Les variantes de la preuve d'enjeu

La preuve d'enjeu possède de nombreuses variantes dont en particulier la preuve d'enjeu par délégation dont le principe est le suivant. Les joueurs engagent de l'argent — comme dans la preuve d'enjeu — pour recevoir des parts de la somme S distribuée périodiquement. Cet argent misé leur donne le droit de participer à l'élection d'une série de délégués qui recevront l'argent distribué et qui dans un réseau de cryptomonnaies joueront le rôle de validateurs. Une redistribution de l'argent reçu par les délégués-validateurs permet à tous les joueurs de profiter de l'argent distribué. Un tel système est intéressant car il limite le nombre de validateurs ; ce qui permet au réseau de fonctionner plus rapidement, et il permet à ceux qui veulent profiter de l'argent distribué de ne pas à avoir à engager des moyens techniques nécessaires pour valider les transactions du réseau et le surveiller.



encore un peu plus du problème réel des cryptomonnaies. Si la distribution du mécène se déroule un très grand nombre de fois, par exemple toutes les heures pendant des années, au lieu de distribuer la somme S en la fractionnant, l'organisateur du jeu pourrait choisir de la distribuer à l'une des enveloppes qui serait choisie au hasard en utilisant la distribution de probabilité fixée par les sommes distribuées avec la méthode de l'enjeu, ou la méthode du travail.

Si, par exemple, la méthode de l'enjeu fixe que l'enveloppe 1 reçoit $\frac{1}{4}$ de S et l'enveloppe 2, reçoit $\frac{3}{4}$ de S (il n'y a que deux enveloppes pour l'exemple), la « méthode de l'enjeu probabiliste » choisit de donner S à l'enveloppe 1 avec une probabilité de $\frac{1}{4}$ et de donner S à l'enveloppe 2 avec une probabilité de $\frac{3}{4}$. L'organisateur jette par exemple un dé tétraédrique et donne S à l'enveloppe 1 si le '1' sort et donne la somme S à l'enveloppe 2 si le '2', le '3' ou le '4' sortent. La répétition de l'utilisation du tirage au sort produit à la longue une attribution de $\frac{1}{4}$ de l'argent distribué au premier joueur, et de $\frac{3}{4}$ au second.

Sur le long terme « la méthode probabiliste de l'enjeu » est donc équivalente à la méthode de l'enjeu. De même, « la méthode probabiliste du travail » est équivalente à la méthode du travail. Les raisonnements proposés plus haut pour les comparer s'appliquent et conduisent à la conclusion que la méthode probabiliste du travail engendre une situation absurde de gâchis économique.

Preuve d'enjeu et preuve de travail

Nous sommes maintenant arrivés exactement au problème qui se pose pour les blockchains anonymes de cryptomonnaies. Le réseau qui fait fonctionner une cryptomonnaie (Bitcoin, Ethereum, Cardano, Solana, etc.) distribue périodiquement une certaine somme d'argent pour qu'il y ait des joueurs, qui, ici, sont les validateurs du réseau, c'est-à-dire ceux qui le font fonctionner. Le réseau les rémunère parce que le réseau fonctionne grâce à eux et que plus ils sont nombreux plus le réseau est décentralisé ; ce qui est souhaitable. La rémunération provient soit de nouvelles unités que le réseau crée, soit des commissions que les utilisateurs paient quand ils utilisent le réseau et qui naturellement doivent aller à ceux qui le font fonctionner. C'est la logique du modèle économique de tels réseaux décentralisés : certains payent pour l'utiliser, d'autres sont rémunérés parce qu'ils le font fonctionner. Dans le cas du Bitcoin, $6,25$ bitcoins sont créés toutes les dix minutes environ et attribués à l'un des validateurs du réseau qui reçoit aussi des commissions provenant des utilisateurs du réseau.

Lorsque les validateurs sont anonymes, ce qui est le cas pour les cryptomonnaies mentionnées, le problème de l'attribution est alors exactement celui du « jeu du mécène et des enveloppes » car les validateurs peuvent apparaître sous plusieurs identités comme les joueurs qui envoient plusieurs enveloppes. En sécurité informatique, la tricherie consistant à apparaître sous plusieurs identités se nomme une « attaque Sybil » (cf. encadré 4). Ces attaques ne sont possibles que sur les réseaux anonymes,

et c'est bien parce que les cryptomonnaies mentionnées sont conçues avec des validateurs anonymes qu'il y a un problème. Si les validateurs étaient clairement identifiés, il serait possible de les rémunérer par la méthode S/N ou sa variante probabiliste. Pour les blockchains où les validateurs sont clairement identifiés, le problème de la distribution de l'incitation n'existe pas ou est très simple. Précisons que l'identification des validateurs, n'oblige pas l'identification des utilisateurs qui peuvent rester anonymes même quand les validateurs sont connus.

Concevoir la méthode de rétribution des validateurs du réseau d'une cryptomonnaies dont les validateurs sont anonymes revient donc exactement à résoudre le « problème du mécène et des enveloppes » décrit plus haut.

La méthode de l'enjeu a été proposée et elle est utilisée par de nombreux réseaux de cryptomonnaies avec des variations dans le détail. Elle prend le nom de « preuve d'enjeu » ou de « preuve de participation ». La méthode du travail prend le nom de « preuve de travail ». Les méthodes adoptées sont les versions probabilistes.

Le travail à faire dans le cas de la preuve de travail n'est pas, bien sûr, la fabrication de rubans en dentelle ensuite détruits. Le travail demandé est la réalisation de calculs dans le but de résoudre un problème combinatoire. Le problème à résoudre est conçu de façon à ce qu'un validateur engageant une puissance de calcul c_i (à mettre en parallèle avec la capacité à fabriquer des rubans chaque semaine) gagnera avec une probabilité proportionnelle à c_i. Ces calculs ne servent à rien d'autre qu'à la distribution de l'incitation et donc dès qu'elle a eu lieu, on ne garde rien des calculs qui sont parfaitement comparables aux rubans détruits.

4. Les attaques Sybil

Lorsqu'un système de communication accepte l'anonymat des utilisateurs, cela engendre des difficultés imprévues. Cela est apparu en informatique à l'occasion de la mise au point des réseaux pair-à-pair, c'est-à-dire sans administrateur central. La possibilité pour une même personne d'utiliser plusieurs pseudonymes différents lui procure des avantages assimilables à une tricherie, comme dans le cas de la distribution par la méthode S/N quand un joueur envoie plusieurs enveloppes. Sur les réseaux sociaux, c'est par exemple un moyen de faire croire que plusieurs personnes soutiennent une même idée alors qu'en réalité il n'y en a qu'une. On dénomme « attaque Sybil » ce procédé.

Le nom provient du titre du livre « Sybil » publié en 1973 par Flora Schreiber qui décrivait le cas d'une malade mentale appelée Sybil dans le livre mais dont la véritable identité est Shirley Mason. Elle était atteinte de trouble dissociatif de l'identité c'est-à-dire de personnalités multiples. Le nom d'attaque Sybil fut introduit en informatique par [9]. Le risque d'attaques Sybil a contraint Satoshi Nakamoto à introduire les preuves de travail dans la cryptomonnaie Bitcoin, bien que la solution des preuves d'enjeu introduite quelques années plus tard soit maintenant reconnue comme bien meilleure, ce que le jeu du mécène rend évident.

Shirley Mason.

Dans le cas de la méthode de la preuve d'enjeu, un validateur engage une somme en la mettant sous séquestre numériquement — ce qu'on sait faire sans avoir à lever l'anonymat — et il gagne avec une probabilité proportionnelle à cette somme qui est l'équivalent de la somme s_i mise dans l'enveloppe envoyée à l'organisateur de la distribution.

L'évolution de ce que nous avons décrit pour « la méthode du travail », s'est produite pour les blockchains qui ont choisi la méthode de « la preuve de travail », ce qui est le cas du Bitcoin. Une industrie du calcul spécialisée s'est progressivement mise en place comparable à l'industrie des rubans de notre histoire. La compétition entre calculateurs est devenue de plus en plus forte et les moyens engagés pour mener les calculs de plus en plus importants. La logique économique incontournable comme dans le cas des rubans a conduit à l'utilisation de moyens industriels de calcul qui font qu'aujourd'hui l'électricité dépensée pour les calculs d'une cryptomonnaie comme le Bitcoin est équivalente à la production de cinq centrales nucléaires au minimum, et sans doute trois fois plus. Cette dépense rogne les revenus des validateurs, comme le coût de fabrication des rubans diminue la somme S réellement distribuée par le mécène. Cette perte est absurde puisque la méthode de la « preuve d'enjeu » l'évite totalement. Il se trouve que le passage d'une méthode à l'autre n'est pas facile. Le Bitcoin ne l'envisage pas et restera économiquement inefficace et écologiquement scandaleux. Ethereum va passer de la preuve de travail à la preuve d'enjeu. Les cryptomonnaies récentes fonctionnent toutes selon la méthode de la preuve d'enjeu ou une variante.

Conclusion

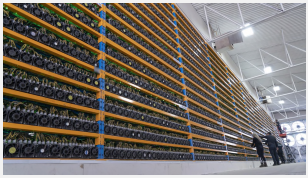
De ces questions concernant le fonctionnement des blockchains et des cryptomonnaies dont « le jeu du mécène et des enveloppes » permet la compréhension, il faut retenir deux points principaux :

— la difficulté initiale provient de l'anonymat des validateurs. Si on y renonce, tout est assez simple. Les blockchains dites privées ou de consortium, qui fonctionnent entre des utilisateurs en nombre limité et connus, ne rencontrent de ce fait aucun problème de dépense électrique ;

— même si on souhaite permettre aux validateurs d'une blockchain de cryptomonnaies de rester anonymes, les solutions de type « preuve d'enjeu » évitent aussi la dépense électrique et l'impact écologique de la « preuve de travail » qui apparaît comme une erreur initiale dans la conception du Bitcoin. Malgré l'admiration qu'on doit à l'invention de la première cryptomonnaie que fut le Bitcoin, sa « preuve de travail » est indubitablement une absurdité petit à petit abandonnée. C'est d'ailleurs indispensable si on souhaite réellement que cette nouvelle sorte de monnaie se développe, permettant l'existence d'un argent liquide numérique anonyme respectueux de la vie privée de chacun.

5. Conséquences des choix faits

Permettre aux validateurs d'une cryptomonnaie de rester anonyme oblige à un système de rémunération délicat alors que le problème est simple quand les validateurs sont bien identifiés, et que personne ne peut se cacher sous plusieurs pseudonymes. Cet anonymat s'il concerne aussi les utilisateurs (ceux qui souhaitent détenir des comptes) ouvre la porte à une multitude de trafics et d'escroqueries. Les plus graves sont les rançongiciels : le pirate installe un programme sur le réseau d'une entreprise, d'une administration ou d'un hôpital ; le programme chiffre les fichiers du système informatique et fait apparaître un message indiquant qu'il faut payer une rançon en cryptomonnaie anonyme pour retrouver les données perdues.



Une usine de minage de la firme Bitfarms au Canada. On aperçoit au fond les ventilateurs de refroidissement qui évacue la chaleur produite par les milliers de machines participant au concours de calcul.

La preuve de travail introduit d'autres escroqueries encore, dont le *crypto-jacking*. Puisque dans un système fonctionnant avec la preuve de travail, c'est la quantité de calculs qu'on est capable de faire qui fixe combien on gagne ; le pirate informatique installe un programme sur des ordinateurs qui ne sont pas à lui et leur demande de faire ces calculs rémunérateurs. Des milliers d'ordinateurs ont ainsi été mis au service de pirates. Cela conduit parfois à des dysfonctionnements graves des systèmes informatiques concernés et à des vols important d'électricité. Il faut remarquer que la preuve d'enjeu n'ouvre pas la possibilité de ce type de vol lié uniquement à la preuve de travail.

Références

[1] Jean-Paul Delahaye. Se libérer du Bitcoin, Introduction aux blockchains et aux cryptomonnaies. Dunod, 2022.

[2] Université de Cambridge. *Bitcoin network power demand*, données mises à jour en continue sur la consommation électrique du réseau Bitcoin, `https://ccaf.io/cbeci` 2022.

[3] Manuel Valente. Qu'est-ce que le *Proof-of-Stake* (preuve d'enjeu) ? `https://www.coinhouse.com/fr/academie/blockchain/proof-of-stake/`. Coinhouse, 2022.

[4] Jean-Paul Delahaye. Au-delà du Bitcoin. Pour la science, n°499, 80-85, Mai, 2019.

[5] Valéria Faure-Muntian, Claude de Ganay, Ronan Le Gleut. Les enjeux technologiques des blockchains. OPECST (Office parlementaire d'évaluation des choix scientifiques et technologiques), `http://www.senat.fr/rap/r17-584/r17-5841.pdf`, 2018.

[6] Jean-Pierre Landau, Alban Genais. Les cryptomonnaies. Rapport au Ministre de l'économie et des finances, `https://www.mindfintech.fr/files/documents/Etudes/Landau_rapport_cryptomonnaies_2018.pdf`, 4 juillet 2018.

[7] Jean-Paul Delahaye. Consommation électrique des cryptomonnaies et des blockchains. Document pour France-Stratégie, « La consommation électrique des technologies disruptives », `http://cristal.univ-lille.fr/~jdelahay/temporaire/DelahayeFranceStrat4juin2018.pdf`, 4 juin 2018.

[8] Jean-Paul Delahaye. Les preuves de travail (Détails sur les concours de calculs utilisés par le Bitcoin). Pour la science, n°438, 80-85, avril 2014.

[9] John Douceur. The Sybil Attack. *International workshop on Peer-To-Peer Systems*, 2002.

Automates cellulaires et robustesse aux erreurs : une perspective mathématique

Irène Marcovici [1]

La difficulté de concevoir des systèmes robustes à des erreurs est une problématique commune à différents domaines de l'informatique. Comment mener à bien un calcul si des erreurs sont susceptibles de se produire ? Comment détecter et corriger des anomalies dans un réseau, en l'absence d'autorité centrale permettant de contrôler l'ensemble du fonctionnement ? Alors que les organismes vivants présentent tous une certaine capacité à se réparer lorsqu'ils sont soumis à une perturbation, c'est rarement le cas des systèmes artificiels, pour lesquels une petite perturbation locale peut mener à un dysfonctionnement complet.

Les automates cellulaires offrent un modèle de choix pour étudier de manière théorique dans quelle mesure un système informatique distribué peut avoir la capacité de se stabiliser en présence d'un bruit aléatoire. Formellement, on suppose que l'information est encodée sur un réseau régulier (ruban infini dans le cas de la dimension 1, grille de dimension 2 ou plus), constitué de *cellules*, qui contiennent chacune un état parmi un ensemble fini d'états possibles. L'action d'un automate cellulaire consiste alors à mettre à jour les états de toutes les cellules simultanément, en appliquant pour chaque cellule une même règle locale de mise à jour, qui dépend seulement de l'état de la cellule et de celui de quelques-unes de ses voisines. Les automates cellulaires constituent ainsi un modèle de calcul parallèle, à la fois suffisamment simple pour permettre une étude mathématique rigoureuse, et suffisamment riche pour apporter un éclairage sur les comportements de systèmes informatiques concrets.

1. Institut Élie Cartan de Lorraine, université de Lorraine.

DOI:10.48556/SIF.1024.19.133 **1024** – Bulletin de la SiF, numéro 19, Avril 2022, pp. 133–136

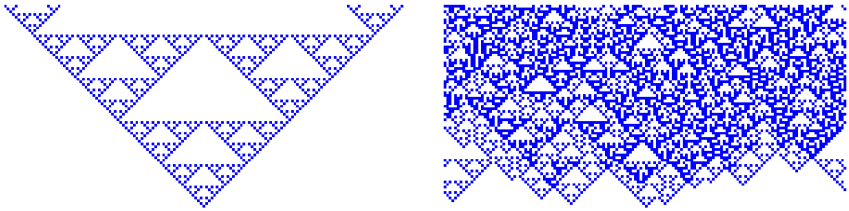


FIGURE 1. Évolution (du bas vers le haut) de l'AC dont la règle locale consiste à calculer la somme (modulo 2) des voisines de gauche et de droite (les carrés bleus représentent les 1, les carrés blancs les 0). À partir d'une configuration initiale constituée d'un unique carré bleu, on observe un triangle de Sierpiński en l'absence d'erreurs *(gauche)*, tandis qu'avec une probabilité $\varepsilon = 0.01$ d'erreurs *(droite)*, l'AC converge vers un état d'équilibre qui ne dépend pas de la configuration initiale.

La figure 1 présente l'évolution d'un automate cellulaire (AC) de dimension 1, à états binaires. Cet exemple élémentaire illustre déjà l'écart entre la simplicité de la règle locale et la complexité qu'elle peut engendrer au niveau macroscopique. Il permet également d'entrevoir l'intérêt d'une perspective mathématique des automates cellulaires.

Calcul bruité

On peut montrer qu'il est possible de simuler l'évolution de n'importe quelle machine de Turing à l'aide d'un automate cellulaire. Mais lorsque des erreurs se produisent, le calcul perd souvent toute signification. Plus précisément, supposons qu'à chaque étape de l'évolution de l'automate cellulaire, chaque cellule est susceptible d'être mise à jour avec un état qui ne correspond pas à la valeur donnée par la règle locale. Même si la probabilité d'une telle erreur est extrêmement petite, au cours de l'évolution, on observe généralement que l'automate cellulaire oublie progressivement toute l'information qui était encodée dans la configuration donnée en entrée, et converge vers un état d'équilibre qui ne dépend plus du tout de la configuration initiale (une illustration en est donnée en figure 1 : en présence d'erreurs, l'automate cellulaire atteint rapidement un comportement *stationnaire*, avec des motifs similaires qui se répètent d'une ligne à l'autre). En termes mathématiques, on dit que le système est *ergodique*. En dimension 1, la célèbre conjecture des taux positifs affirmait que tout automate cellulaire soumis à des erreurs aléatoires est ergodique. Peter Gàcs [Gàcs01] a réfuté cette conjecture en 2001, en proposant un contre-exemple extrêmement sophistiqué, dans un article de plus de 200 pages. Mais si on se limite aux familles d'automates cellulaires les plus simples, nous avons montré récemment

dans un travail avec Mathieu Sablik et Siamak Taati [MST19] qu'ils ne sont effectivement pas robustes à des erreurs aléatoires. En dimension supérieure ou égale à 2, des familles de contre-exemples avaient été proposées par Andrei Toom [Toom80] dès 1980, mais le problème n'est pas trivial pour autant, et encore aujourd'hui, l'ergodicité des automates cellulaires bruités soulève de nombreuses questions, qui intéressent aussi bien les informaticiennes et les informaticiens que les spécialistes de probabilités et de systèmes dynamiques.

Auto-correction de pavages

Avec Nazim Fatès et Siamak Taati [FMT22], nous avons fait un pas de côté pour étudier la capacité des automates cellulaires à corriger des erreurs dans des pavages. En un sens, il s'agit de comprendre comment reproduire sur un modèle artificiel simple certains mécanismes d'auto-stabilisation présents dans la nature.

On se donne un pavage de la grille, c'est-à-dire un coloriage des cellules de la grille bi-dimensionnelle avec une palette finie de couleurs, où l'agencement des couleurs doit respecter certaines contraintes locales. On suppose que les couleurs de certaines cellules sont arbitrairement changées par de nouvelles couleurs, ne respectant pas forcément les contraintes du pavage. Peut-on retrouver une configuration respectant les contraintes du pavage, par une procédure locale ? En d'autres termes, existe t-il un automate cellulaire, qui, à partir de n'importe quelle perturbation finie du pavage, se stabilise au bout d'un temps fini sur une configuration valide ? Même si par essence, les pavages sont définis par un ensemble de contraintes locales, il n'est pas toujours aisé d'élaborer des règles locales d'auto-correction.

Pour nous forger une intuition, considérons le cas des k-coloriages. Un k-coloriage est un coloriage des cellules de la grille avec une palette contenant k couleurs, qui doit respecter la contrainte que deux cellules adjacentes (c'est-à-dire qui ont une arête commune) doivent toujours être de couleurs différentes. Pour $k = 2$, les seules configurations valides correspondent aux deux configurations en damier, et on peut utiliser leur périodicité spatiale pour construire un automate cellulaire permettant de reconstruire rapidement le damier si des erreurs ont été introduites. Pour $k \geqslant 4$, on peut au contraire exploiter le fait que le nombre de couleurs disponibles est suffisamment grand pour offrir une certaine souplesse, ce qui permet également de retrouver rapidement une configuration valide par des opérations locales. Le cas $k = 3$ est quant à lui plus complexe. Nous n'avons pas pu trouver de règle auto-correctrice à ce jour et nous pensons que le problème est difficile. En effet, comme illustré en figure 2, il existe des configurations où seules deux cellules adjacentes ont la même couleur et où, pourtant, il faut modifier une zone arbitrairement grande pour retrouver une configuration valide, ce qui laisse penser que si une solution existe, elle ne saurait être simple.

Si le statut des 3-coloriages reste ouvert à ce jour, on sait cependant que tous les pavages ne peuvent pas être corrigés facilement par un automate cellulaire. Nous

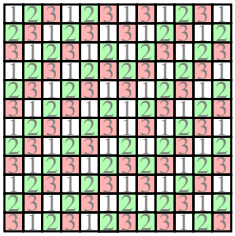


FIGURE 2. Exemple de configuration d'un 3-coloriage où seules deux cellules adjacentes (au milieu) sont de même couleur blanche, mais pour laquelle il est nécessaire de modifier une zone beaucoup plus grande si l'on souhaite retrouver une configuration valide.

avons en effet montré que si P≠NP, il existe des pavages pour lesquels la correction par un automate cellulaire, si tant est qu'elle soit possible, ne peut pas être effectuée en temps polynomial par rapport à la taille de la zone qui a été altérée.

À l'inverse, nous avons présenté différentes familles de pavages pour lesquelles des mécanismes d'auto-stabilisation efficaces peuvent être proposés, typiquement en temps linéaire par rapport à la taille de la modification. Dans ce cas, on obtient même un résultat plus fort : l'automate cellulaire permet non seulement de corriger des erreurs qui auraient été introduites sur une zone finie de la configuration, mais aussi de corriger des erreurs aléatoires réparties sur toute la grille, si la probabilité d'erreur est suffisamment petite.

Notre approche, à la croisée entre mathématiques et informatique, nous a fait étudier le problème de la robustesse aux erreurs dans le cadre abstrait d'un réseau régulier sur lequel l'évolution des différentes cellules est régie par une même règle locale. Les modèles retenus, malgré leur simplicité, apportent un éclairage sur le fonctionnement des réseaux distribués actuels, en fournissant quelques pistes pour développer leurs capacités d'auto-régulation, et en permettant également de mieux comprendre les limites théoriques qui existent.

Références

[Gàcs01] P. Gács. Reliable cellular automata with self-organization. *J. Stat. Phys.*, 103(1–2) : 45–267, 2001.

[MST19] I. Marcovici, M. Sablik, et S. Taati. Ergodicity of some classes of cellular automata subject to noise. *Electron. J. Probab.*, 24 : 1–44, 2019.

[FMT22] N. Fatès, I. Marcovici, et S. Taati. Self-stabilisation of cellular automata on tilings. *Fundam. Inform.*, 2022 (à paraître).

[Toom80] A. Toom. Stable and attractive trajectories in multicomponent systems. *Multicomponent random systems, Adv. Probab. relat. Top.*, 6 : 549–575, 1980.

Reproductibilité numérique : enjeux de crédibilité pour les expériences de simulation

Paul-Antoine Bisgambiglia [1] et David R.C. Hill

Cet article vise à sensibiliser les utilisateurs d'outils de simulation et à compléter les éléments produits pendant la journée de la Société informatique de France consacrée à la reproductibilité en 2021. Nous proposons de revenir sur les définitions de base, sur la distinction entre répétabilité et reproductibilité numérique et nous apportons un éclairage sur les pratiques permettant de consolider la crédibilité des résultats de simulation informatique.

Introduction

Karl Popper a profondément impacté la production des connaissances dans de nombreux domaines [21], il reste encore aujourd'hui une référence majeure en épistémologie. Les dérives récentes des pratiques scientifiques ont conduit la société à réaliser l'importance des critères de scientificité qu'il avait mis en avant, notamment un critère majeur qui est la reproductibilité d'expériences scientifiques.

Nous pouvions lire, dans un article du journal Le Monde, daté du 2 octobre 2017, qui a pour titre « *La rigueur scientifique à l'épreuve de la reproductibilité* » que : « *dès 2005, John Ioannidis, de l'université Stanford, suggérait de façon provocatrice dans un article de PloS Medicine :* « la plupart des résultats scientifiques sont faux », car impossibles à reproduire.

1. Maître de conférences à l'université de Corse.

DOI:10.48556/SIF.1024.19.137

Dans [11], nous pouvons lire que « *même les estimations les plus prudentes de la recherche en biomédecine placent le taux de reproductibilité à moins de 50 %.* ». Au même moment en 2019, le Guardian faisait un bilan des travaux financés en Grande-Bretagne par le fond pour l'excellence scientifique, et même ici, le taux de reproductibilité pour les travaux scientifiques du domaine médical tombait à 11 %. Dans le secteur du numérique, nous pourrions nous attendre à de bien meilleures « performances » car nos machines et nos piles logicielles sont supposées toujours déterministes, mais une étude poussée [3], datant de la même époque a montré que la reproductibilité des travaux de recherche en informatique ne dépassait guère les 30 %.

Une recherche [2] rapide sur *Web of Science* avec les termes *reproducible research* donne 51 862 résultats, et avec le terme *reproducibility* donne 139 857 résultats. Le domaine qui semble produire le plus d'études en ce sens est la chimie analytique avec 22,565 résultats, soit plus de 16 %. La production par an est croissante, elle a presque doublé en 10 ans passant de 5000 à plus de 9000 articles entre 2012 et 2021. Le français est la quatrième langue de production, mais ne représente que 0.5 % des articles alors que plus de 6 % de ces articles sont produits par des équipes françaises. Ce focus montre l'intérêt croissant du domaine. Nous retrouvons également trois articles significatifs, dont un éditorial, sur le sujet dans la revue Nature fin 2021 [1, 10, 19]. Nous pouvons y lire que reproduire les résultats scientifiques est difficile et chronophage, mais crucial, ainsi l'effort doit être partagé entre les auteurs, les laboratoires et la communauté scientifique : « *L'ensemble de la communauté scientifique doit reconnaître que la réplication d'une observation ou d'un résultat est indispensable, elle permet d'acquérir une assurance essentielle au progrès de la science : elle montre qu'une observation ou un résultat est suffisamment solide pour encourager de futurs travaux* [3]. ».

L'idée première derrière la notion de reproductibilité en informatique est que l'on puisse reproduire une expérience qui a été partagée au sein de la communauté scientifique dans le but d'obtenir strictement les mêmes résultats. Dans la partie qui suit nous aborderons des définitions en se rappelant des pratiques dans les sciences expérimentales plus anciennes, comme la biologie et la physique. Tout le déroulé d'une expérimentation est noté sur un cahier de laboratoire et partagé avec les pairs afin d'en permettre la reproductibilité et d'en valider les conclusions, ces approches sont transposables aux expériences numériques de simulation.

Dans son exposé pour la Société informatique de France, Christophe Pouzat, définit la recherche reproductible comme : « *une démarche qui consiste à fournir aux*

2. Recherche effectuée le 26 Novembre 2021.

3. Traduction de « *The entire scientific community must recognize that replication is not for replication's sake, but to gain an assurance central to the progress of science : that an observation or result is sturdy enough to spur future work* ».

lecteurs d'articles, d'ouvrages, etc., l'ensemble des données et des programmes accompagnés d'une description algorithmique de la façon dont les programmes ont été appliqués aux données pour obtenir les résultats présentés. ». Obtenir les résultats présentés est bien souvent problématique, ce phénomène est accentué en informatique par la prolifération de codes de calcul ou de simulation, faciles à utiliser et mis à disposition sans contexte d'utilisation ou avec un manque de documentation. Cependant, une raison très fréquente qui empêche d'obtenir les résultats produits dans une étude publiée est clairement l'absence d'accès au code informatique. La simulation informatique s'est développée au point de devenir un outil précieux de production de connaissances et d'aide à la décision, et c'est même le seul outil, indispensable, pour l'exploration des systèmes dits complexes. De nombreux processus de décision utilisent des résultats de simulation informatique pour guider les décideurs : aérodynamisme de voiture, structure de bâtiment, propagation de virus, etc. Dans bien des cas, les outils ou codes de simulation qui ont fait d'énormes progrès en termes d'ergonomie et de facilité d'utilisation. Il arrive alors qu'ils soient utilisés sans assez de recul sur les mécanismes sous-jacents : méthodes de résolution ou de discrétisation, générateurs de nombres pseudo-aléatoires, ordonnanceurs d'événements, mécanismes de gestion de la concurrence, mécanismes de compilation, etc. La pile logicielle qui permet d'utiliser une application est supposée toujours maîtrisée, ce qui n'est pas toujours le cas pour des applications de calcul souvent sophistiquées, et également dans bien des chaînes logicielles modernes ou il n'est plus « évident » d'avoir deux logiciels identiques lors de deux *builds* successifs [18].

Les causes de non-reproductibilité sont nombreuses et complexes : développement trop rapide de codes de simulation personnels, complexité des codes, difficultés et coûts de maintenance des codes existants, mauvais usage des générateurs de nombres pseudo-aléatoires ou tout simplement manque de rigueur, ou encore et surtout de temps, etc.

Ces éléments sont abordés dans [2], nous les avons traités dans le cadre de la thèse de [12] qui s'intéressait à un aspect facilitant la reproductibilité des expériences de simulation, à savoir la formalisation du processus de modélisation et l'usage de code de simulation univoque. Nos propositions avaient pour but de faciliter la reproductibilité de simulations multi-agents, mais peuvent compliquer l'usage du paradigme agent et ne concernent qu'une partie limitée de la problématique de la science reproductible. Nous avions pris soin des aspects stochastiques des simulations depuis 2014 dans [8], puis nous nous sommes intéressés aux problèmes de reproductibilité des simulations stochastiques distribuées dans [15, 17]. Une rapide association de ces différents travaux est proposée dans [13]. Forts de ces différentes expériences, nous proposons dans cet article de revenir sur les définitions du domaine et proposons de lister et résumer quelques bonnes pratiques.

Définitions

Dans le cadre d'expériences numériques, [7, 6, 17, 11] proposent un état de l'art et des définitions issues de la littérature. Deux nuances sont dégagées dans ces études sur la reproductibilité : (1) la reproduction de l'expérience numérique et (2) la reproduction exacte des résultats numériques obtenus. Cette reproductibilité numérique — parfois appelée *bitwise reproducibility* en anglais — est nécessaire ne serait-ce que pour mettre au point les programmes numériques sur ordinateur (sans débogage, plus de logiciels).

La nuance entre reproductibilité et répétabilité est également mise en avant : la répétabilité consiste à retrouver les mêmes résultats lorsque deux expériences sont menées avec les mêmes paramètres d'entrées, avec des matériels, des méthodes et des contextes identiques. On parle aussi en anglais de *run to run reproducibuility*. À grande échelle sur des supercalculateurs, il arrive que cette répétabilité attendue soit perdue après le lancement de deux ou plusieurs expériences de simulations en tout point identiques. L'utilisation d'optimisations au sein des nouveaux microprocesseurs (*out of order execution*) ou des nouvelles instructions machines fusionnées du type FMA(*Fused Mutiply Add*) ou de type vectorielles SIMD (*Single Instruction Multiple Data*) sont des causes de non-reproductibilité qui deviennent fréquentes [16]. Il devient alors difficile de mettre au point les programmes à grande échelle que l'on fait tourner sur des clusters plus lents que les supercalculateurs en production sur lesquels les erreurs de répétabilité ont été constatées.

La reproductibilité se veut donc plus générale que la répétabilité. Parmi les apports de la reproductibilité, [7] nous rappellent qu'elle « constitue une méthode et un standard pour juger de la pertinence d'une expérience numérique publiée et donc des conclusions qui en découlent ». Entre la notion de répétabilité et de reproductibilité, différents degrés sont suggérés dans la littérature. La reproductibilité implique des changements, mais l'obtention de la même conclusion scientifique. C'est un critère essentiel pour la scientificité d'une étude selon Karl Popper. Des équipes différentes, des méthodes différentes, des instruments différents donnent (heureusement) les mêmes conclusions scientifiques. Dans la phrase précédente, le terme « conclusion scientifique » est préféré au terme « résultat ». Dans un contexte numérique, il faut dans bien des cas complexes savoir se contenter de résultats « seulement » similaires, donnant la même conclusion scientifique. C'est un critère de validation, etc.

Obtenir les mêmes résultats numériques lorsque quelque chose a changé dans le contexte d'exécution revient à parler de portabilité. Le standard de calcul numérique flottant IEEE 754 a été élaboré à ce propos. Nous pouvons aussi citer l'importance du respect de l'ordre d'exécution des calculs flottants au sein des compilateurs. Nous sommes proches de la répétabilité par le fait que le même résultat numérique est

obtenu — se répète —, mais comme le contexte a changé nous sommes dans le cadre large de la reproductibilité dite *bitwise reproducibility*.

Pour asseoir les définitions, il faut se replacer au niveau épistémologique — dans ce cadre la répétabilité suppose que nous n'avons aucun changement dans la réalisation de l'expérience. La répétabilité n'est donc pas un élément de la démarche scientifique au sens de Karl Popper [20].

Chris Drumond rappelle même que ce n'est pas ce que l'on attend dans une démarche scientifique [9]. Mais il se trouve que cette caractéristique est essentielle et attendue dans le contexte de développements informatiques purement déterministes. Sans cette propriété comment mettre au point des programmes, comment les déboguer [17]. Dans [5], nous trouvons quatre niveaux de reproductibilité : (1) la disponibilité du code source, (2) les aspects non-déterministes du calcul, (3) la similitude du plan d'expérience, et (4) les sorties du modèle.

Dans leurs travaux, les auteurs de [22] définissent également plusieurs niveaux de reproductibilité :

(1) l'examen par les pairs, qui est la méthode traditionnelle de publication (les travaux et résultats décrits sont jugés crédibles par la communauté scientifique) ;

(2) la recherche réplicable, où les outils permettant de reproduire les mêmes résultats sont fournis ;

(3) la recherche vérifiable, où les mêmes conclusions peuvent être atteintes indépendamment du code source fourni par l'auteur ;

(4) la recherche validable, où suffisamment de ressources (sources et données) sont archivées afin de permettre de défendre les résultats fournis ;

(5) la recherche ouverte, où tous les éléments utilisés pour arriver aux résultats présentés sont fournis en accès libre et documentés.

Il est important de préciser le cas des calculs stochastiques. Soit la source de hasard utilisée est maîtrisée : générateurs pseudo-aléatoires ou quasi-aléatoires et dans ce cas, nous avons un modèle maîtrisé (et déterministe !) du hasard et nous sommes à même de répéter numériquement les expériences des simulations stochastiques. Soit, nous avons une source non déterministe telle que celle utilisée pour du calcul quantique (avec un algorithme et des circuits quantiques) et dans ce cas nous attendons la reproductibilité des conclusions scientifiques, mais nous n'aurons jamais de répétabilité numérique.

Dans les travaux cités, nous voyons aussi apparaître les notions de vérification et de validation. Ce sont des notions importantes pour la communauté des chercheurs en simulation et plus largement en ingénierie logicielle. Elles sont employées dans cette communauté avec un sens précis et sont les piliers de la crédibilité que l'on peut accorder aux études de simulation. Le travail de vérification vise à montrer qu'un modèle informatisé (opérationnel et exécutable) représente bien un modèle

conceptuel dans des limites de précisions spécifiées. En ce qui concerne la validation des modèles, il s'agit de montrer qu'un modèle opérationnel informatisé possède une plage de précision satisfaisante compatible avec l'application prévue du modèle dans son domaine d'applicabilité (son cadre expérimental). Pour plus de détails sur les différentes techniques de vérification et de validation, le lecteur intéressé pourra consulter le chapitre VII de l'ouvrage suivant [4].

C'est en se souciant de ces aspects tout au long de la conception et de l'utilisation des simulations que notre domaine scientifique pourra gagner en crédibilité.

Discussion

Les conclusions établies par [5, 22, 7, 17] se rejoignent, et nous pouvons les résumer de la manière suivante : pour permettre un degré de reproductibilité se rapprochant de la « recherche ouverte », un changement dans la culture de publication est nécessaire. Ce changement doit être impulsé non seulement par les auteurs, mais également par les éditeurs. Il est recommandé aux premiers de fournir les éléments permettant de reproduire l'expérience de simulation (données, code source, etc.), et aux seconds d'encourager cet effort durant le processus de publication en offrant les supports numériques d'archivage et de diffusion de ce contenu.

Des efforts sont faits, par exemple dans le formulaire d'examen des articles des conférences SCS est apparu en 2020 un item d'évaluation portant sur la reproductibilité des résultats de simulation, c'est également le cas maintenant pour de nombreux journaux. Il y a également de nombreuses initiatives de la communauté comme un MOOC [4] d'Inria « recherche reproductible : principes méthodologiques pour une science transparente » qui met en avant l'utilisation d'outils tels que :

— markdown pour la prise de notes structurées ;

— des outils d'indexation (DocFetcher et ExifTool) ;

— gitlab / github pour le suivi de version et le travail collaboratif ;

— notebooks (Jupyter, Rstudio ou Org-mode) ;

— ou encore, la journée de la SIF [5].

Il convient de citer également : le projet Software Heritage d'Inria pour l'archivage de code source ; les ReproHackathon du GDR MaDICS qui visent à tester les capacités des systèmes de workflows disponibles à reproduire une expérience scientifique, etc.

Ce processus d'échange des méthodes, codes et protocoles d'expériences augmente la confiance dans les résultats. Il est donc en partie garant du respect et de

4. https://www.fun-mooc.fr/courses/course-v1:inria+41016+self-paced/about.
5. https://www.societe-informatique-de-france.fr/journee-reproductibilite.

la crédibilité de la méthode scientifique en cette période où les controverses scientifiques montrent plus souvent les limites de la méthode et l'impact des conflits d'intérêts sur la production de résultats qui peuvent être mis à jour grâce à l'absence de reproductibilité.

Dans son dernier ouvrage [14], Stephen Grossberg, pionnier des modèles cognitifs utilisés aujourd'hui dans l'apprentissage profond, précise que nous ne pouvons pas nous fier de façon aveugle à ces techniques. Les principales raisons sont d'une part, la non explicabilité, et d'autre part son inadéquation à plusieurs types de domaines applicatifs où la technique peut engendrer des « oublis catastrophiques ». Nous constatons également que lorsque ces algorithmes fonctionnent, il peut être impossible de comprendre pleinement pourquoi et que lors d'apprentissages sur de nouvelles bases de données, des souvenirs construits précédemment disparaissent de façon arbitraire.

Si la reproductibilité des algorithmes est assez simple à obtenir (la classification a cette souplesse), la possibilité de rejouer ou de répéter strictement les expériences est nécessaire pour progresser dans la connaissance (dans la science). Restons prudent dans l'usage de ces techniques pour toutes applications où les conséquences peuvent être lourdes – notamment dans le cadre d'applications médicales par exemple. Il est fréquent que ces modèles utilisent des simulations stochastiques en nombre massif pour la phase d'apprentissage. C'était notamment le cas pour l'entrainement d'Alpha Go qui avait battu Lee Sedol le champion du monde de Go. Pour éviter les écueils de ces nouvelles méthodes, Stephen Grossberg propose d'approfondir les travaux en lien avec la logique floue à partir d'un modèle qu'il appelle ART (*Adaptive Resonance Theory*). Ce type de modèle permettrait d'éviter les problèmes évoqués et on peut espérer valider des modèles de ce type couplés aux simulations, pour explorer et entraîner les modèles.

Références

[1] O. B. Amaral and K. Neves. Reproducibility : expect less of the scientific paper. 597, 2021/09.

[2] P.-A. Bisgambiglia. *Habilitation à diriger des recherches : Les expériences virtuelles de simulation comme outils d'aide à la prise de décisions des données au processus de décision*. PhD thesis, Université de Corse, Université de Corse, 07 2021.

[3] C. Collberg and T. A. Proebsting. Repeatability in computer systems research. *Communications of the ACM*, 59(3) :62–69, 2016.

[4] P. Coquillard and D. R. C. Hill. *Modélisation et simulation d'écosystèmes : Des modèles déterministes aux simulations à événements discrets*. Masson, Recherche en Écologie, 1997. ISBN 2-225-85363-0.

[5] O. Dalle. On reproducibility and traceability of simulations. In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, pages 1–12, Dec. 2012.

[6] V. T. Dao. *Calcul à haute performance et simulations stochastiques. Etude de la reproductibilité numérique sur architectures multicore et manycore*. PhD thesis, LIMOS – UMR CNRS 6158, Université Clermont Auvergne, Clermont, Mar. 2017. 00000.

[7] V. T. Dao, V. Breton, H. Nguyen, and D. R. C. Hill. La reproductibilité des simulations stochastiques parallèles et distribuées utilisant le calcul à haute performance. *Journées DEVS Francophone*, pages 109–117, 2016.

[8] V. T. Dao, L. Maigne, V. Breton, H. Nguyen, and D. R. C. Hill. Numerical reproducibility, portability and performance of modern pseudo random number generators : Preliminary study for parallel stochastic simulations using hybrid xeon phi computing processors. In *European Simulation And Modelling Conference*, pages 80–87, 2014.

[9] C. Drummond. Replicability is not reproducibility : nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning workshop 26th International Conference for Machine Learning 2009*, 2009.

[10] N. Editorials. Replicating scientific results is tough — but crucial. 600, 2021/12.

[11] B. G. Fitzpatrick. Issues in reproducible simulation research. *Bulletin of mathematical biology*, 81(1) :1–6, 2019.

[12] R. Franceschini. *Approche formelle pour la modélisation et la simulation à évènements discrets de systèmes multi-agents*. phdthesis, Université de Corse Pasquale Paoli, Dec. 2017.

[13] R. Franceschini, P.-A. Bisgambiglia, and D. R. C. Hill. Reproducibility Study of a PDEVS Model Application to Fire Spreading. In *Proceedings of the 50th Computer Simulation Conference*, SummerSim '18, pages 29 :1–29 :11, San Diego, CA, USA, 2018. Society for Computer Simulation International.

[14] S. Grossberg. *Conscious Mind, Resonant Brain : How Each Brain Makes a Mind*. Oxford University Press, 2021.

[15] D. R. C. Hill. Parallel Random Numbers, Simulation, and Reproducible Research. *Computing in Science & Engineering*, 17(4) :66–71, July 2015. 00001.

[16] D. R. C. Hill. Numerical reproducibility of parallel and distributed stochastic simulation using high-performance computing. In *Computational Frameworks*, pages 95–109. Elsevier, 2017.

[17] D. R. C. Hill, V. T. DAO, C. Mazel, and V. Breton. Reproductibilité et répétabilité numérique. constats, conseils et bonnes pratiques pour le cas des simulations stochastiques parallèles et distribuées. *Technique et Science Informatiques*, 36(3-6) :243, 2017.

[18] C. Lamb and S. Zacchiroli. Reproducible builds : Increasing the integrity of software supply chains. *IEEE Software*, 2021.

[19] N. C. Nelson et al. Understand the real reasons reproducibility reform fails. *Nature*, 600(7888) :191–191, 2021.

[20] A. O'Hear. Karl popper : Philosophy and problems. *Royal Institute of Philosophy*, 1995.

[21] K. Popper. *Logique de la découverte scientifique*. édition originale : Logic der Forschung, Springer, Wien, 1934, payot edition, 1973.

[22] V. Stodden, J. Borwein, and D. H. Bailey. Setting the default to reproducible. *computational science research. SIAM News*, 46(5) :4–6, 2013. 00030.

L'apprentissage non-supervisé et ses contradictions

Jérémie Sublime [1, 2]

L'apprentissage machine, également appelé apprentissage artificiel [2], ou encore *machine learning* en anglais, est un domaine scientifique lié à l'intelligence artificielle et qui se trouve à la frontière entre l'informatique et les statistiques. L'objectif de ce domaine est la conception et l'étude d'algorithmes capables « d'apprendre » à partir de données. Ces algorithmes d'apprentissage sont ensuite utilisés pour des tâches diverses et variées : analyse et traitement d'images (i.e. reconnaissance de visages), détection d'anomalies (i.e. connexions frauduleuses), systèmes de recommandation (Google, Youtube, Amazon, Netflix, etc.), tri et catégorisation d'information, et même pour apprendre à des programmes à jouer aux échecs, au Go [3], ou aux jeux vidéos [4].

On divise généralement l'apprentissage machine en trois grandes catégories : l'apprentissage supervisé, non-supervisé et par renforcement.

Dans l'apprentissage supervisé, les algorithmes sont entraînés à reconnaître et catégoriser des données à partir d'exemples étiquetés. On apprendra, par exemple, à un réseau de neurones à faire la différence entre des photos de chats ou de chiens en lui montrant des centaines d'images et en lui précisant à chaque fois de quel animal il s'agit de sorte qu'il finira par apprendre ce qui caractérise un chat ou un chien, et saura reconnaître ces caractéristiques sur une photo qu'il n'aura jamais vue.

1. ISEP, École d'ingénieurs du numérique.
2. LIPN - CNRS UMR 7030, LaMSN - Université Sorbonne Paris Nord.
3. https://www.bbc.com/news/technology-35785875.
4. DeepMind/AlphaStar.

DOI:10.48556/SIF.1024.19.145 **1024** – Bulletin de la SiF, numéro 19, Avril 2022, pp. 145–156

Dans l'apprentissage non-supervisé, conçu à la base comme une tâche exploratoire, les algorithmes sont amenés à traiter des données non-étiquetées pour y trouver des structures, des groupes d'objets similaires, ou des anomalies. Si on reprend l'exemple précédent avec les photos de chats et de chiens, on fournira cette fois-ci toutes les photos à l'algorithme sans lui préciser de quel animal il s'agit. Avec un peu de chance, la méthode non-supervisée s'apercevra toute seule qu'il y a deux animaux différents selon les photos, et fera un tri pour séparer les chiens et les chats. Mais il se peut aussi qu'elle décide de les trier par couleur (en mélangeant chiens et chats), qu'elle fasse de multiples catégories par race, ou encore qu'elle fasse un tri selon des critères difficilement compréhensibles pour un humain (la couleur du papier peint par exemple) en ne se préoccupant nullement des animaux présents sur les images. On parle de *clustering* lorsque l'apprentissage non-supervisé regroupe ensemble des données considérées comme similaires selon des critères plus ou moins transparents. Les groupes ainsi formés sont appelés *clusters*.

L'apprentissage par renforcement est un domaine un peu différent des précédents puisqu'on cherche ici à influencer le comportement d'un algorithme d'intelligence artificielle à partir de récompenses si l'algorithme se comporte comme on le souhaite, et de pénalités dans le cas contraire. Ce type d'apprentissage est davantage utilisé pour apprendre aux algorithmes à jouer à des jeux (en choisissant les bons coups à jouer) plutôt que pour traiter des données.

L'apprentissage supervisé est sans doute la catégorie la plus connue des chercheurs tous domaines confondus, avec ses réseaux de neurones aux performances extraordinaires en analyse d'image. L'apprentissage par renforcement, quant à lui, est le plus médiatisé auprès du grand public pour ses succès lorsque les machines battent les meilleurs humains sur des jeux ou des tâches que l'on estimait il y a quelques années encore trop complexes pour qu'un algorithme puisse vaincre un jour un humain. Quant à moi, je travaille depuis le début de ma carrière de chercheur sur l'apprentissage non-supervisé, la branche moins connue de l'apprentissage, la plus mal posée aussi, et qui sera le sujet de cet article. J'aborderai un certain nombre de contradictions et de questions que l'on peut voir émerger dans ce domaine un peu particulier.

Un apprentissage non-supervisé détourné de son but exploratoire initial ?

Lorsqu'il est présenté à des étudiants dans l'enseignement supérieur, l'apprentissage non-supervisé est généralement abordé sous l'angle du *clustering* : cette tâche exploratoire que j'ai évoquée plus haut et qui consiste à regrouper ensemble des données similaires dans des *clusters*. Ce cours sur le *clustering* fait généralement partie d'un module plus large d'apprentissage machine et est l'occasion de présenter les principales familles d'algorithmes de *clustering* : on présentera les K-moyennes [11]

pour les méthodes dites « basées centroïdes », le *clustering* basé sur la densité des données dans l'espace [5, 1], le *clustering* hiérarchique [16] qui construit les clusters sur plusieurs niveaux comme un dendrogramme, et parfois la version statisticienne des K-moyennes sous la forme des modèles de mixtures gaussien. Pour chaque famille d'algorithmes, on détaillera les grands principes, la complexité algorithmique, mais aussi quels types de clusters peuvent et ne peuvent pas être trouvés par ces algorithmes. Les structures que chacun des algorithmes peuvent identifier sont généralement illustrées comme dans la figure 1 ci-après (qui évoque aussi les temps de calcul). En effet, chaque famille d'algorithmes de *clustering* présuppose des formes de clusters à trouver et leur séparabilité.

C'est ce dernier point qui vient généralement mettre le premier coup de canif au concept de l'apprentissage non-supervisé comme méthode exploratoire : l'idée de devoir donner à l'avance presque toutes les informations sur les clusters qu'on va chercher si on veut avoir une petite chance de les trouver.

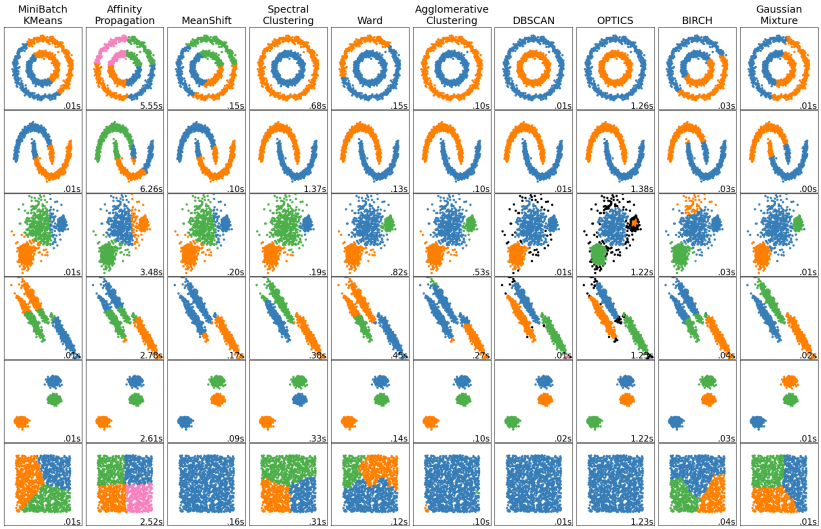


FIGURE 1. Quelques exemples de méthodes de clustering appliquées à diverses configurations de points dans l'espace. © Scikit-learn [5]

Certains réfutent parfois ce point en expliquant qu'il s'agit — avec le choix de l'algorithme — de faire une hypothèse a priori sur cette forme, et non pas de la

5. https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py.

connaître à l'avance. Cet argument est à mon avis naïf, étant donné l'impossibilité d'émettre une hypothèse éclairée sans avoir déjà une très bonne idée de ce qu'on cherche. Il souligne cependant bien le côté « problème mal posé » du clustering.

Si on regarde maintenant les applications concrètes de l'apprentissage non-supervisé dans la vie d'un chercheur ou d'un industriel, on s'aperçoit assez vite que dans la plupart des cas pratiques ce n'est pas tellement le côté exploratoire qui motive son utilisation, mais plutôt l'absence de données annotées. En effet, l'immense majorité des tâches de traitement de données ou d'images qui utilisent des techniques d'intelligence artificielle dans la vie réelle, sont des tâches pour lesquelles on cherche des classes précises et connues : dans mon exemple d'introduction, on voulait séparer les chiens et les chats, en imagerie médicale, on voudra par exemple classifier si des tissus sont cancéreux ou non, en imagerie satellite on voudra identifier et classer différents types de bâtiments vus du ciel, en marketing digital on voudra identifier des catégories sociaux-économiques bien précises pour cibler des publicités en ligne, etc. Cependant, il se trouve que ces domaines sont parfois confrontés à une absence de données annotées (absence totale ou nombre insuffisant) pré-existantes pour une application précise, empêchant alors d'avoir recours aux classiques algorithmes d'apprentissage supervisés. Dans ces conditions, l'utilisation de l'apprentissage non-supervisé — qui lui ne nécessite pas de données annotées — peut sembler être une alternative séduisante par rapport au coût de création d'une base suffisante de données étiquetées. On se retrouve alors de fait avec une utilisation détournée de l'apprentissage non-supervisé, qui a perdu tout caractère exploratoire, et est utilisé par défaut pour pallier un manque de données étiquetées. Et ce détournement n'est pas sans conséquences puisque ces algorithmes n'ont absolument pas été conçus pour que leurs clusters convergent comme par magie vers les classes recherchées pour les applications réelles. Les résultats sont souvent assez décevants, surtout lorsqu'ils sont comparés aux performances qu'on aurait pu avoir avec des étiquettes et un bon algorithme supervisé.

Mais alors pourquoi continue-t-on de détourner les méthodes non-supervisées plutôt que d'étiqueter des données pour passer ensuite au supervisé ? On peut citer deux explications complémentaires : la première est la relative simplicité des données et des méthodes, jusqu'à la fin des années 90. Cela permettait aux méthodes non-supervisées de conserver des performances correctes dans le meilleur des cas, et de pouvoir envisager un étiquetage manuel pas trop coûteux en cas de nécessité de passer à du supervisé. Dans le même temps, la complexité des données a également explosé, creusant ainsi les écarts de performances entre les algorithmes. La seconde explication est la percée récente des réseaux de neurones profonds qui sont particulièrement gloutons en données étiquetées pour pouvoir atteindre les excellentes performances que nous leur connaissons. Si je résume, nous sommes donc dans une conjoncture où les données sont à la fois plus nombreuses, plus complexes et donc

plus difficile à étiqueter, et nous avons aussi des algorithmes supervisés qui nécessitent de plus en plus de données annotées.

L'imagerie satellite et l'imagerie médicale — deux domaines que je connais bien — sont de bons exemples de l'apparition de plus en plus fréquente de ce type de problème. En imagerie satellite, la variété des paysages, des capteurs (dont la résolution et le nombre de bandes varient) et des applications, fait que des données annotées pour une application seront rarement réutilisables pour un autre problème avec des images d'une autre région. En imagerie médicale, on retrouve ce même problème de format d'acquisition qui change et de variété des applications (rendant les annotations déjà faites pas forcément utiles d'un problème à l'autre), mais aussi des conflits d'experts qui ne sont pas d'accord sur la façon dont certaines images complexes devraient être annotées, ajoutant donc en plus un problème de confiance dans les annotations. Ainsi, lorsque ces experts métier (géographes, professeurs de médecine, ou autres) se tournent vers les spécialistes de l'apprentissage supervisé pour voir si on ne pourrait pas automatiser l'interprétation de leurs images, ou même prédire de futures évolutions avec le dernier réseau de neurones à la mode, on leur répond souvent que c'est possible mais qu'il faudra fournir des centaines (parfois des milliers) d'images précisément annotées. Si les très gros instituts et conglomérats peuvent peut-être fournir ces annotations (et parfois les partager par la suite avec le reste de la communauté), les autres doivent renoncer à l'idée ou tenter leur chance avec un spécialiste de l'apprentissage non-supervisé.

L'apprentissage non-supervisé a fait beaucoup de progrès depuis ses débuts, et est lui aussi entré dans l'ère des réseaux de neurones. Il existe donc quelques réponses dans l'arsenal non-supervisé qui permettent de s'attaquer à des images ou des données complexes avec des outils adaptés : l'auto-encodeur [9, 7] reste à mon sens l'architecture de base la plus utilisée lorsqu'on veut combiner apprentissage non-supervisé et réseaux de neurones. Cette architecture qui est présentée dans sa forme la plus simple sur la Figure 2 ci-après, vise à compresser et à reconstruire les données en entrées via un encodeur et un décodeur symétriques qui pourront être composés de diverses couches de neurones qui dépendront de l'application. L'entonnoir de l'auto-encodeur permet généralement de récupérer une version des données débruitée et encodée dans un espace latent qui sera plus facilement utilisable, par exemple par une méthode de clustering classique. Ainsi, l'auto-encodeur permet à la fois de réutiliser la majorité des composants (couches et fonctions d'activation) des réseaux de neurones classiques, et peut-être combiné soit à des méthodes de clustering classiques, soit à quelques couches neuronales supplémentaires spécifiques pour obtenir l'algorithme non-supervisé voulu.

Il n'en reste pas moins que malgré la possibilité technique de construire des réseaux de neurones non-supervisés capables de s'attaquer à des données aussi complexes que des images satellites ou médicales, nous restons dans ce cas de figure, pas du tout exploratoire, où l'on attend d'une méthode non-supervisée qu'elle trouve des

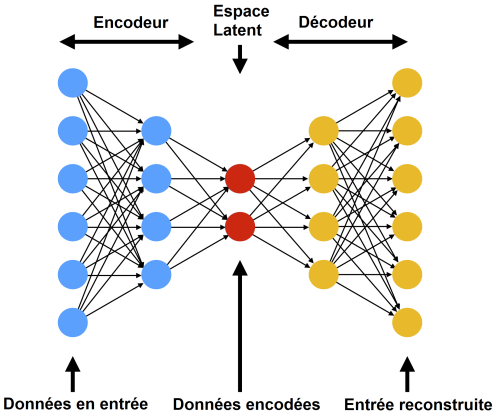


FIGURE 2. Exemple d'une architecture de base d'auto-encodeur qu'on pourra modifier selon l'application.

clusters qui vont coller avec des classes réelles très précises sur lesquelles les experts eux-même ne sont parfois pas d'accord. Pour prendre un exemple issu de mon manuscrit d'habilitation [15] : "*Attendre d'un réseau de neurones non-supervisé qu'il vous trouve des lésions sur une image médicale, c'est comme demander à un enfant de 5 ans de colorier ce qui lui* ***paraît remarquable*** *sur cette même image. Les deux ont à peu près les mêmes capacités visuelles, le réseau de neurones sera peut-être légèrement meilleur en coloriage, et aucun des deux n'a la moindre idée de ce qui est sur l'image ou de ce que vous attendez de lui. Vous aurez votre coloriage, mais il n'est pas certain que le résultat soit le résultat attendu.*"

Dans ces conditions, et puisque nous assumons que nous ne sommes plus dans un contexte exploratoire mais bien dans un cas où l'on sait ce qu'on cherche, où l'on a juste pas assez de données annotées : puisque l'on possède quelques données annotées, pourquoi ne pas les utiliser pour aider l'algorithme de clustering ? Ce cadre a d'ailleurs été étudié et s'appelle l'apprentissage semi-supervisé. Hélas, les choses ne sont pas si simples. Si effectivement l'apprentissage semi-supervisé est bien un domaine qui existe et qui consiste à fournir des données étiquetées et d'autres non-étiquetées à un algorithme d'apprentissage pour qu'il apprenne mieux, pour le moment, il concerne principalement l'amélioration des performances d'algorithmes supervisés auxquels il faudra quand même fournir un maximum de données étiquetées, et pour lesquels fournir quelques données non-étiquetées pourra améliorer les performances en fin d'apprentissage. En revanche, les expériences pour faire l'inverse — avoir majoritairement des données non-étiquetées et fournir quelques données étiquetées à un algorithme non-supervisé (comment ?) — sont encore rares et très balbutiantes. On aura donc tendance à considérer que le semi-supervisé n'existe

pas vraiment, ou reste en tout cas très supervisé.

De cette section, on pourrait conclure que non seulement les méthodes non-supervisées sont finalement assez peu utilisées dans un cadre vraiment exploratoire, mais surtout que même quand on n'a pas d'autre choix que de les utiliser il y a de fortes chances que les résultats soient décevants. Cependant, nous allons voir par la suite qu'on peut tout de même utiliser les algorithmes non-supervisés avec une certaine efficacité, à condition de bien appréhender les limites de la non-supervision.

L'apprentissage non-supervisé serait-il quand même un peu supervisé ?

Étant donné ce qui a été évoqué précédemment sur les difficultés des méthodes non-supervisées à répondre aux attentes de trouver des classes réelles sur des applications pratiques, je me suis souvent posé la question de savoir comment d'autres chercheurs et moi-même arrivions tout de même à atteindre des résultats honorables (pour du non-supervisé) sur des problèmes complexes avec nos algorithmes. En effet, si on regarde la littérature scientifique, cet apprentissage non-supervisé — qui semble ne rien avoir pour lui — reste utilisé par de nombreux chercheurs dans de nombreux domaines, en particulier à cause du manque de données étiquetées. Et les résultats sont parfois impressionnants.

C'est un récent co-encadrement de stage et une rétrospection sur mon expérience personnelle qui m'ont apporté quelques éléments de réponse que je vais partager avec vous. Le stage en question avait pour but de faire de la segmentation de lésions liées à la DMLA[6] sur des images médicales en infrarouge [14] : un problème assez classique d'imagerie médicale donc, avec des lésions plus ou moins complexes à détourer selon les images, des experts pas d'accord sur les cas complexes, et pas assez d'images bien annotées pour entraîner un réseau de neurones supervisé de type UNet [12]. Notre stagiaire c'est donc tourné vers les WNets [17], le réseau de neurones équivalent aux UNets en non-supervisé pour faire de la segmentation. Ce réseau repose d'ailleurs en partie sur l'architecture d'auto-encodeur évoquée précédemment. Les premiers résultats furent ce qu'on pouvait attendre : mauvais. Les lésions qui nous intéressaient n'étaient que rarement séparées du reste, coupées en de multiples classes, ou complètement ignorées au profit d'autres éléments structurels présents dans les images et qui avaient sans doute été jugés plus remarquables par l'algorithme non-supervisé. Malgré tout, nous avons persévéré : notre stagiaire avait pour habitude de nous montrer de nouveaux résultats tous les 3 ou 4 jours au fil des modifications apportées sur les fonctions d'activation des neurones, les changements de fonctions de perte, ou de tous autres paramètres de son algorithme. Il s'est avéré qu'au fil des modifications et de nombreux essais et erreurs, les résultats se sont

6. Dégénérence maculaire liée à l'âge, une pathologie de l'œil.

lentement mais sûrement améliorés. Il m'est alors apparu comme assez évident que faute de pouvoir superviser un algorithme directement et automatiquement avec de grands volumes de données étiquetées, il était possible de le guider au fil des essais à partir de résultats visuels et des résultats sur les quelques données étiquetées à disposition : il est possible de choisir certains paramètres en fonction de ce qui fonctionne ou de ce qui ne fonctionne pas. Mais n'y a-t-il pas là une forme de supervision qui ne dit pas son nom ?

En effet, en y réfléchissant bien, le processus que nous avons décrit est plus ou moins le même que celui que j'ai pu constater chez la majorité des chercheurs qui font de l'apprentissage non-supervisé : on part d'une méthode de base (et nous avons vu que le choix de cette méthode va favoriser certaines formes de clusters et peut donc déjà être vu comme une forme de supervision), puis on va modifier et guider cette méthode par tous les moyens détournés possibles et imaginables jusqu'à avoir des résultats satisfaisants sur l'application qui nous intéresse. Et ce processus de guidage — quel que soit sa forme [6] — nécessite une certaine expertise et demande du temps ; en particulier par rapport aux méthodes supervisées pour lesquelles il suffit généralement d'adapter un peu les formats en entrée de l'algorithme pour que de bons résultats sortent assez rapidement. Je pense donc pouvoir affirmer que le secret d'un apprentissage non-supervisé qui donne de bons résultats est que tout le monde « triche » en compensant l'absence de supervision via des données étiquetées par une autre forme de supervision : celle d'un processus de guidage manuel de l'algorithme (du choix de la méthode de base aux paramètres) jusqu'à atteindre les résultats voulus.

Cette reconnaissance d'une supervision autrement que directement par les données en non-supervisé (a priori courante, mais rarement admise ouvertement) soulève tout de même quelques questions :

— peut-on toujours parler d'apprentissage non-supervisé dans ce contexte ?

— pourrait-on automatiser ce processus ? En effet, si on regarde de plus près, cela ressemble à de l'apprentissage par renforcement qui serait fait à la main. Mais, c'est également assez proche de l'autoML [8] tel qu'on le pratique en apprentissage supervisé ;

— quelles conséquences le recours à ce processus peut-il avoir lors de la comparaison des performances de plusieurs méthodes non-supervisées ?

Je vais m'attarder sur la dernière question qui est assez essentielle pour de nombreux chercheurs en apprentissage : nous sommes tous plus ou moins soumis à une certaine injonction à publier, et chacun sait que — à tort ou à raison — pour être publiable un algorithme applicatif doit bien souvent montrer qu'il fait mieux (ou au moins aussi bien) que ses concurrents de la littérature. Dès lors qu'on se place dans le prisme d'une application réelle, on peut oublier les indices classiques de clustering [13, 3], et on se concentrera sur une évaluation par le prisme des indices supervisés

classiques tels que la précision, le rappel ou le score F1. Bien que nous soyons en contexte non-supervisé ici, nous avons tout de même ces quelques images annotées à notre disposition qui, si elles ne sont pas assez nombreuses pour entraîner une méthode supervisée, suffisent généralement pour avoir une évaluation quantitative. Or, si on admet l'existence de ce processus de supervision détournée des méthodes non-supervisées pour les rendre plus compétitives, on s'aperçoit assez vite que la comparaison ne sera pas équitable entre une méthode nouvellement proposée par des chercheurs qui l'auront sur-optimisée et qui ne pourra que battre ses concurrents non-supervisés réutilisés depuis d'autres applications plus ou moins proches et qui eux n'auront pas forcément reçu le même niveau d'optimisation par une supervision détournée. Bien que n'ayant pas de solution à ce problème, je pense que ce sujet méritait d'être détaillé, d'autant que ce phénomène du nouvel algorithme qui écrase complètement des concurrents pourtant a priori sérieux (mais sans doute pris sur l'étagère) est assez courant dans la littérature non-supervisée.

A l'issue de cette seconde partie de réflexion, on s'aperçoit finalement que la clé d'un apprentissage non-supervisé qui donne des résultats corrects sur des données complexes est très certainement d'apporter la supervision autrement que par les étiquettes, de manière à tout de même guider l'algorithme. Si ce processus est aujourd'hui à mon avis massivement pratiqué, il reste très artisanal et serait à perfectionner. Est-ce à dire pour autant qu'en faisant cela, les méthodes non-supervisées peuvent faire aussi bien que les méthodes supervisées ? La réponse courte sera non. Certes cette supervision détournée améliore considérablement les résultats, mais elle

ne battra jamais une supervision directe et automatisée par les étiquettes des classes recherchées. Cependant, le gain de temps reste considérable par rapport au processus de construction à la main d'un corpus de données annotées de quelques milliers d'images.

Conclusion

Nous avons abordé dans ce texte plusieurs aspects et contradictions de l'apprentissage non-supervisé. Nous avons vu que contrairement à l'idée que l'on s'en fait et à la façon dont on l'enseigne, il n'est *a priori* que rarement exploratoire : on l'applique comme solution par défaut sur des problèmes bien supervisés pour lesquels on n'a pas ou peu de données étiquetées. En y regardant de plus près, on s'aperçoit aussi que ce qu'on appelle apprentissage non-supervisé est finalement assez souvent supervisé de manière détournée : que ce soit par le choix de l'algorithme qui va favoriser certains clusters constituant ainsi une première forme de supervision, ou bien dans des choix plus pointus d'architecture et de modèles qui seront des points encore plus forts de supervision.

Personnellement, je n'aime pas le qualificatif de « non-supervisé » qui est à mon avis trompeur. Je ne suis pas le seul. Certains préfèrent maintenant parler de *self-supervised learning* [18, 4, 10] : l'apprentissage des données par les données. C'est un meilleur nom. Et cette idée fonctionne très bien sur certaines tâches. Cependant, et bien que ce ne soit pas le sujet de ce texte, on pourrait argumenter que les principaux succès du *self-supervised learning* ont été accomplis en indiquant aux algorithmes quelle partie des données pourrait être apprise à partir de quelle autre, et donc avec une supervision externe. Je reste par conséquent peu convaincu que les données seules puissent jamais apporter la supervision nécessaire pour les nombreuses tâches de classification ou de traitement d'image avancées qu'on souhaite réaliser sans avoir à annoter des milliers de données manuellement. En effet, ces annotations manuelles servent à indiquer à l'algorithme quel est son but et quelles classes il doit trouver : elles servent à lui expliquer pourquoi il est intéressant de grouper par exemple des chiens et des chats plutôt que de grouper des photos majoritairement vertes. Aussi, il me semble que penser l'apprentissage non-supervisé sans concevoir qu'il y a un but derrière est illusoire. Une forme de supervision — peu importe laquelle, et qui pourra être plus ou moins complexe selon le but recherché — doit exister pour guider l'algorithme.

Au delà de la reconnaissance des forces et des limites de l'apprentissage non-supervisé, et d'un éventuel changement de nom pour mieux coller à la réalité de son utilisation, il me semble que les prochaines avancées pour sortir l'apprentissage non-supervisé de ses contradictions devraient se concentrer sur deux points évoqués dans cet article :

(1) le développement de méthodes *faiblement supervisées* où le guidage d'un algorithme vers des classes réelles pourrait se faire à partir de très peu d'exemples annotés en plus de données non-annotées ;

(2) la reconnaissance et la formalisation de l'actuel processus de supervision détournée, de manière à pouvoir l'utiliser plus efficacement et à assurer des comparaisons équitables de méthodes non-supervisées.

Références

[1] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander. OPTICS : Ordering Points To Identify the Clustering Structure. In *In ACM SIGMOD international conference on Management of data*, pages 49–60. ACM Press, 1999.

[2] A. Cornuéjols and L. Miclet. *Apprentissage artificiel : Concepts et algorithmes*. Eyrolles, June 2010.

[3] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2) :224–227, Feb. 1979.

[4] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2070–2079, 2017.

[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[6] P. Gançarski, T.-B.-H. Dao, B. Crémilleux, G. Forestier, and T. Lampert. Constrained Clustering : Current and New Trends. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of AI Research*, volume 2. Springer, 2020.

[7] X. Guo, X. Liu, E. Zhu, and J. Yin. Deep Clustering with Convolutional Autoencoders. In D. Liu, S. Xie, Y. Li, D. Zhao, and E.-S. M. El-Alfy, editors, *Neural Information Processing*, pages 373–382. Springer International Publishing, 2017.

[8] I. Guyon, I. Chaabane, H. J. Escalante, S. Escalera, D. Jajetic, J. R. Lloyd, N. Macià, B. Ray, L. Romaszko, M. Sebag, A. R. Statnikov, S. Treguer, and E. Viegas. A brief review of the chalearn automl challenge : Any-time any-dataset learning without human intervention. In F. Hutter, L. Kotthoff, and J. Vanschoren, editors, *Proceedings of the 2016 Workshop on Automatic Machine Learning, AutoML 2016, co-located with 33rd International Conference on Machine Learning (ICML 2016), New York City, NY, USA, June 24, 2016*, volume 64 of *JMLR Workshop and Conference Proceedings*, pages 21–30. JMLR.org, 2016.

[9] G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786) :504–507, 2006.

[10] Y. LeCun. Self-supervised learning. In *Plenary talk at the 8th International Conference on Learning Representations, ICLR 2020, Online*, 2020.

[11] J. B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[12] O. Ronneberger, P. Fischer, and T. Brox. U-Net : Convolutional Networks for Biomedical Image Segmentation. volume 9351, pages 234–241, 10 2015.

[13] R. Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics.*, 20 :53–65, 1987.

[14] C. Royer, J. Sublime, F. Rossant, and M. Pâques. Unsupervised approaches for the segmentation of dry ARMD lesions in eye fundus cslo images. *J. Imaging*, 7(8) :143, 2021.

[15] J. Sublime. *Contributions to modern unsupervised learning : Case studies of multi-view clustering and unsupervised Deep Learning. (Contributions à l'apprentissage non-supervisé moderne : Applications aux cas du clustering multi-vue et de l'apprentissage profond non-supervisé)*. 2021.

[16] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301) :236–244, 1963.

[17] X. Xia and B. Kulis. W-net : A deep model for fully unsupervised image segmentation. *CoRR*, abs/1711.08506, 2017.

[18] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer. S4l : Self-supervised semi-supervised learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1476–1485, 2019.

Récurrence nœthérienne pour le raisonnement de premier ordre

Sorin Stratulat [1]

La récurrence nœthérienne est un des principes les plus généraux de raisonnement formel. Dans le cadre du raisonnement de premier ordre, nous présentons une classification de ses instances pouvant être partagées en instances basées sur des termes et des formules. Nous donnons un aperçu du raisonnement par récurrence nœthérienne basée sur des termes et sur des formules, puis nous établissons des relations entre eux. Enfin, nous présentons une méthodologie pour la certification du raisonnement basé sur des formules à l'aide de l'assistant de preuve Coq.

Introduction

Le raisonnement par récurrence est bien adapté pour valider des propriétés sur des structures de données non-bornées, comme les naturels et les listes, ainsi que des spécifications basées sur des fonctions récursives (protocoles, algorithmes distribués, etc). Sa force réside dans la possibilité d'utiliser des formules dont on n'a pas besoin de montrer la validité, en tant qu'*hypothèses de récurrence*.

Un des principes les plus généraux du raisonnement par récurrence, la *récurrence nœthérienne*, utilise des *ordres nœthériens* (ou bien fondés) pour garantir l'application correcte des hypothèses de récurrence.

1. Université de Lorraine, CNRS, LORIA.

DOI:10.48556/SIF.1024.19.157 **1024** – Bulletin de la SiF, numéro 19, Avril 2022, pp. 157–170

Définition 3 (le principe de récurrence nœthérienne). *Étant donné un ensemble (potentiellement infini) $\mathscr{E}$ muni d'un ordre nœthérien $<$, le principe de récurrence nœthérienne peut être formalisé par la formule d'ordre supérieur suivante :*

$$\forall \varphi, (\forall m \in \mathscr{E}, (\forall k \in \mathscr{E}, k < m \Rightarrow \varphi(k)) \Rightarrow \varphi(m)) \Rightarrow \forall p \in \mathscr{E}, \varphi(p), \tag{1}$$

où φ est un symbole de prédicat quantifié universellement.

Le principe de récurrence nœthérienne permet de vérifier si une propriété définie sur les éléments de l'ensemble $\mathscr{E}$ est satisfaite par tous les éléments de $\mathscr{E}$. Ainsi, pour une propriété φ et un élément $m \in \mathscr{E}$ donnés, toute formule $\varphi(k)$, avec $k < m$, peut être utilisée dans la preuve de $\varphi(m)$, en tant qu'hypothèse de récurrence. L'argument de correction est donné par la propriété de l'ordre nœthérien $<$ qui interdit toute suite *infinie* strictement décroissante d'éléments de $\mathscr{E}$. Ce principe se démontre simplement par contradiction si on suppose qu'il y a un élément $x_0 \in \mathscr{E}$ qui ne satisfait pas φ, alors il existe par contraposition un élément plus petit $x_1 \in \mathscr{E}$ qui ne satisfait pas φ. On peut raisonner sur x_1 comme on avait fait pour x_0 pour déduire qu'il y a un élément x_2 de $\mathscr{E}$ plus petit que x_1 et qui ne satisfait pas non plus φ, etc. On conclut ainsi qu'il y a une suite $\ldots < x_2 < x_1 < x_0$ strictement décroissante infinie d'éléments de $\mathscr{E}$ qui ne satisfont pas φ. Ceci contredit le fait que $<$ est un ordre nœthérien.

Nous nous sommes intéressés à l'application du principe de récurrence nœthérienne dans le raisonnement de premier ordre. Ceci se résume à produire des instances de premier ordre de la formule (1). Dans le reste de l'article, nous allons d'abord classifier ces instances selon une taxonomie que nous avons proposée pour la première fois dans [35] et qui prend en considération les cas où les éléments de $\mathscr{E}$ sont soit des termes, soit des formules de premier ordre. Ensuite, nous établissons des relations entre ces deux situations, puis nous faisons un état de l'art des principes de récurrence nœthérienne basés sur des termes et sur des formules. Enfin, nous montrons comment formaliser et certifier le raisonnement basé sur des formules à l'aide de l'assistant de preuve Coq [42], en résumant la méthodologie présentée dans [37].

Instances basées sur des termes et sur des formules de premier ordre

Comme exemple illustratif, on va prouver par récurrence la conjecture

$$x + 0 = x, \text{ pour tout naturel } x \tag{2}$$

à partir des axiomes

$$0 + x = x, \text{ pour tout naturel } x \tag{3}$$

$$S(x) + y = S(x + y), \text{ pour tout naturel } x \text{ et } y \tag{4}$$

qui définissent l'addition '+' sur les naturels à l'aide des symboles de constructeurs 0 et S (la fonction 'successeur'). '=' est le seul symbole de prédicat et les formules sont des *égalités* de la forme $s = t$.

Durant les preuves, on va employer un raisonnement dit *équationnel* qui implémente le principe de Leibniz de remplacements des égaux par des égaux [2] en utilisant les parties gauche et droite des égalités. Les variables sont implicitement quantifiées universellement.

On va prouver $n+0=n$ pour un naturel n arbitraire qui peut être soit 0, soit le successeur d'un autre naturel n'. Dans le premier cas, $0+0=0$ est une instance de l'axiome (3) lorsque x est 0. Dans le deuxième cas, $S(n')+0=S(n')$ est équivalent à $S(n'+0)=S(n')$ si on remplace la partie gauche par la partie droite de l'égalité $S(n')+0=S(n'+0)$, qui est une instance de (4) lorsque x est n' et y est 0. L'hypothèse de récurrence $n'+0=n'$ permet de remplacer $S(n'+0)$ par $S(n')$ afin d'obtenir l'identité $S(n')=S(n')$ et de finir la preuve.

L'usage correct de $n'+0=n'$ peut être justifié par deux instances différentes du principe de récurrence nœthérienne, lorsque les éléments de l'ensemble $\mathscr{E}$ de la Definition 3 sont i) des (vecteurs de) termes, ou ii) des formules de premier ordre. Pour faire la distinction entre les ordres employés dans les deux cas, on va dénoter $<$ par $<_t$ pour le premier cas, et par $<_f$ pour le deuxième cas.

Récurrence nœthérienne basée sur des termes (RNT). Lorsque $\mathscr{E}$ est un ensemble de vecteurs de termes, (1) devient une formule de premier ordre si φ est fixé à un symbole de prédicat de premier ordre donné. L'instance de (1) obtenue ainsi est :

$$(5) \qquad (\forall \overline{m} \in \mathscr{E}, (\forall \overline{k} \in \mathscr{E}, \overline{k} <_t \overline{m} \Rightarrow \varphi(\overline{k})) \Rightarrow \varphi(\overline{m})) \Rightarrow \forall \overline{p} \in \mathscr{E}, \varphi(\overline{p})$$

où les symboles supralignés représentent des vecteurs de termes.

Un exemple de principe de RNT est celui de *Peano* : pour prouver une formule $P(x)$ pour tout naturel x, il est suffisant de la prouver pour x étant 0, puis $S(x')$, où x' est une nouvelle variable naturelle. Dans le deuxième cas, le principe de Peano permet d'utiliser $P(x')$ en tant qu'hypothèse de récurrence parce que $x' <_t S(x')$ lorsqu'on prend $<_t$ comme étant la relation « plus petit que » sur les naturels.

Exemple 1 (La preuve de $x+0=x$ par récurrence de Peano)**.** *Dans notre exemple, on définit $P(x)$ en tant que l'égalité $x+0=x$. L'hypothèse de récurrence fournie par le principe de Peano dans la preuve de $P(S(n'))$ est $P(n')$, i.e., $n'+0=n'$. C'est ce qui a été demandé.*

D'autre part, $n'+0=n'$ peut être aussi définie comme hypothèse de récurrence légitimée par des principes de récurrence nœthérienne basés sur des formules, comme c'est expliqué ci-dessous.

Récurrence nœthérienne basée sur des formules (RNF). Cette fois-ci, les éléments de $\mathscr{E}$ de (1) sont des formules de premier ordre. Dans ce cas, φ est un symbole de prédicat de second ordre puisqu'il a des formules de premier ordre comme arguments. Pour que (1) devienne une formule de premier ordre, une solution serait de définir φ comme étant la relation d'identité de second ordre telle que $\varphi(x) = x$, pour toute formule de premier ordre $x \in \mathscr{E}$. L'instance de la formule (1) est dans ce cas :

$$(\forall m \in \mathscr{E}, (\forall k \in \mathscr{E}, k <_f m \Rightarrow k) \Rightarrow m) \Rightarrow \forall p \in \mathscr{E}, p \tag{6}$$

Le principe de RNF ainsi obtenu dit qu'on a le droit d'utiliser une formule k dans la preuve d'une autre formule m si k est plus petite que m.

Exemple 2. *Le principe de RNF peut s'appliquer lorsque la preuve $x+0=x$ a été générée en utilisant un ordre $<_f$ qui respecte les conditions suivantes*[2] *:*

- $x+0=x <_f S(x+0)=S(x)$*, pour tout naturel x,*
- $S(x+0)=S(x) <_f S(x)+0=S(x)$*, pour tout naturel x, et*
- $<_f$ *est nœthérien et* stable par substitutions*, i.e., si $s <_f t$ alors $s\sigma <_f t\sigma$, pour toute substitution σ qui remplace des variables par des termes.*

On définit $\mathscr{E}$ comme étant l'ensemble de toutes les formules produites pendant la preuve, i.e., $\{n+0=n, 0+0=0, S(n')+0=S(n'), S(n'+0)=S(n'), S(n')=S(n')\}$. Dans la preuve de $S(n')+0=S(n')$, on a le droit d'utiliser en tant qu'hypothèse de récurrence toute formule de $\mathscr{E}$ plus petite. Dans notre cas, ce serait $n'+0=n'$, i.e., un renommage de la première formule pour lequel $n'+0=n' <_f S(n'+0) = S(n') <_f S(n')+0=S(n')$.

Théorème 1. *La relation d'identité de second ordre est la plus générale instanciation de φ qui permettrait de faire de (1) une formule de premier ordre.*

Démonstration. Soient $\mathscr{E} = \{\varphi_1, \ldots, \varphi_n, \ldots\}$ un ensemble de formules de premier ordre, p un symbole de prédicat de second ordre et $p(\varphi)$ une formule de premier ordre pour tout $\varphi \in \mathscr{E}$, et $<_p$ un ordre nœthérien tels que :

$$\forall \varphi_i \in \mathscr{E}, (\forall \varphi_j \in \mathscr{E}, \varphi_j <_p \varphi_i \Rightarrow p(\varphi_j)) \Rightarrow p(\varphi_i)) \Rightarrow \forall \varphi \in \mathscr{E}, p(\varphi) \tag{7}$$

On définit

- le nouvel ensemble de formules $\mathscr{E}' = \{p(\varphi) \mid \varphi \in \mathscr{E}\}$, et
- l'ordre $<_{p'}$ tel que $p(\varphi_i) <_{p'} p(\varphi_j)$ si $\varphi_i <_p \varphi_j$, pour toutes $\varphi_i, \varphi_j \in \mathscr{E}$.

On peut remarquer que $<_{p'}$ est nœthérien et que la formule (7) implique

$$\forall \varphi_i \in \mathscr{E}', (\forall \varphi_j \in \mathscr{E}', \varphi_j <_{p'} \varphi_i \Rightarrow \varphi_j) \Rightarrow \varphi_i) \Rightarrow \forall \varphi \in \mathscr{E}', \varphi$$

qui est une instance de (6) lorsque $\mathscr{E}$ est $\mathscr{E}'$ et $<_f$ est $<_{p'}$. □

2. Un tel ordre peut être défini sur des égalités en tant qu'extension multi-ensemble d'un ordre multi-ensemble sur les chemins basé sur la précédence croissante sur les symboles de fonction 0, S, et $+$ [2].

Principes de RNT

Les principes de RNT définissent les hypothèses de récurrence de manière *explicite* dans le cadre des *schémas de récurrence* [1], souvent issus de l'analyse des fonctions et des structures de données récursives. Étant donné une formule φ, un schéma de récurrence identifie d'abord un sous-ensemble de variables de φ à instancier, appelées des *variables de récurrence*, puis définit des instances de φ en termes d'hypothèses de récurrence et de *conclusions de récurrence* en spécifiant les hypothèses qui peuvent être utilisées dans la preuve de chaque conclusion. L'ensemble de conclusions doit être choisi de manière que sa validité implique celle de φ.

Un exemple de principe de récurrence basé sur des schémas de récurrences, issus des spécifications sortées et basées sur des constructeurs, est la *récurrence structurelle* [28], qui généralise le principe de Peano ainsi que les principes de récurrence mathématiques usuels.

Définition 4 (récurrence structurelle)**.** *Soit φ une formule à vérifier pour tout élément d'une sorte S à l'aide d'un schéma de récurrence à n conclusions. Le principe de la récurrence structurelle est défini par la formule :*

$$(\bigwedge_{i=1}^{n}(\forall x_1, \ldots, x_{n_i}, \varphi(x_{i_1}) \wedge \ldots \wedge \varphi(x_{i_k}) \Rightarrow \varphi(f_i(x_1, \ldots, x_{n_i}))) \Rightarrow \forall x, \varphi(x)$$

où les variables $x_{i_1}, \ldots, x_{i_k}$ sont celles parmi $x_1, \ldots, x_{n_i}$ qui ont la sorte S et toute fonction f_i ($i \in [1..n]$) est d'arité n_i et de codomaine S.

L'ordre de récurrence est implicite mais il existe en tant que relation « plus petite que » sur les naturels représentant la profondeur des arbres syntaxiques associés aux termes donnés comme arguments à φ.

A son tour, la récurrence structurelle est généralisée par la *récurrence sur des ensembles couvrants* [44] qui a été inspirée par l'idée proposée dans [8] selon laquelle les schémas de récurrence sont issus des définitions récursives des fonctions qui apparaissent dans la conjecture à prouver. La récurrence sur des ensembles couvrants suppose que chaque sorte S est caractérisée, ou *couverte*, par un ensemble fini de termes de sorte S appelé *ensemble couvrant* (de termes). La notion d'ensemble couvrant peut être généralisée par un ensemble de vecteurs de termes qui couvrent un produit de sortes $S_1 \times S_2 \times \ldots$

Définition 5 (récurrence sur des ensembles couvrants)**.** *Soient Ψ un ensemble couvrant non-vide $\{\overline{t_1}, \ldots, \overline{t_m}\}$ de vecteurs de termes de sorte $\mathscr{S}$ et φ une formule à vérifier pour tout élément de $\mathscr{S}$. Le principe de la récurrence sur des ensembles couvrants est défini par la formule :*

$$(\bigwedge_{i=1}^{m}(\bigwedge_{j=1}^{m_i} \forall \overline{k_i^j} \in \mathscr{S}, \overline{k_i^j} <_t \overline{t_i} \Rightarrow \varphi(\overline{k_i^j})) \Rightarrow \varphi(\overline{t_i})) \Rightarrow \forall \overline{t} \in \mathscr{S}, \varphi(\overline{t})$$

Les inconvénients les plus importants des principes de récurrence basés sur des schémas sont liés à la gestion des hypothèses de récurrence, lorsqu'elles sont i) définies mais inutilisées, et ii) nécessaires mais non incluses dans le schéma ou impossibles à produire par la méthode basée sur l'analyse de la récursivité. D'autre part, son avantage principal réside dans la définition locale de l'ordre de récurrence, qui se passe au niveau du schéma de récurrence. Ceci permet plus de flexibilité dans la gestion des ordres pendant le déroulement d'une preuve ; les contraintes d'ordre sont vérifiées une seule fois, au moment de la définition des schémas de récurrence. De plus, le raisonnement par récurrence basé sur des schémas peut être facilement intégré dans les *systèmes d'inférence* des démonstrateurs en termes de *règles d'inférence* correctes.

Un autre inconvénient serait présenté par la difficulté de traiter naturellement la *récurrence mutuelle*, adaptée au raisonnement sur des spécifications basées sur des définitions mutuellement récursives des fonctions ou des structures de données. Ceci est dû au fait que les hypothèses et les conclusions des schémas de récurrence sont des instances d'une *même* formule. Cependant, quelques solutions partielles ont été proposées pour palier cet inconvénient. Boyer et Moore [9] construisent des schémas de récurrence à partir d'une nouvelle fonction qui appelle les différentes fonctions définies mutuellement en fonction de la valeur d'un argument supplémentaire. Dans d'autres circonstances, la conjecture doit être généralisée ou des lemmes auxiliaires sont rajoutés par les utilisateurs pour spécifier des propriétés cruciales des fonctions mutuellement définies. Une solution plus automatique a été proposée par Kapur et Subramaniam [23] pour gérer une classe de fonctions mutuellement récursives par leur transformation dans des fonctions (simplement) récursives. L'idée est de déplier les définitions des fonctions lorsque les arguments sont des ensembles couvrants afin d'arriver aux appels simplement récursifs. Une autre idée serait de proposer des schémas de récurrence multi-prédicats [7] en associant un prédicat à chacune des fonctions mutuellement récursives, ce qui élimine le besoin de dépliage de fonctions. Cependant, si la conjecture intègre plusieurs symboles de fonction récursive, les schémas de récurrence doivent être combinés [8].

Principes de RNF

La technique de *preuve par récurrence sans récurrence*, aussi connue sous le nom de *preuve par cohérence*, est la première qui intègre une implémentation du principe de RNF. Proposée par Musser [29], elle utilise l'*algorithme de complétion* de Knuth-Bendix [25] pour prouver des propriétés inductives. La méthode peut prouver qu'un ensemble d'égalités est une conséquence d'un ensemble cohérent d'axiomes égalitaires orientables en règles de réécriture par i) leur ajout aux axiomes, ii) leur orientation dans des règles de réécriture, et iii) par leur cohérence avec les axiomes lorsque le processus de complétion *sature*, i.e, aucune nouvelle égalité n'est générée par rapport à un certain critère de redondance. Le processus de saturation demande que la

stratégie de preuve soit *équitable* afin de garantir le traitement ultérieur de toute nouvelle conjecture. Avec le temps, la méthode a été améliorée [20, 16, 17, 21, 22, 3, 27]. Pour un aperçu sur cette technique de preuve, le lecteur peut consulter [12, 13].

Sans utiliser la saturation, la *récurrence implicite* est une technique de preuve qui sépare les axiomes des conjectures à prouver. Comme la technique de preuve par cohérence, elle est basée sur la réécriture, ce qui permet de construire des règles d'inférence *réductives* qui remplacent à chaque étape de preuve une conjecture φ par un ensemble potentiellement vide de nouvelles conjectures qui sont soit trivialement vraies (e.g. tautologies), soit plus petites par rapport à (des instances de) φ. Dans ce dernier cas, les règles de réécriture utilisées doivent être issues soit des axiomes soit des autres conjectures plus petites que φ. Au début de son évolution, Reddy [31] a proposé la méthode appelée *la récurrence par réécriture de termes*, et mis au point une procédure qui calcule des *ensembles couvrants de formules*, obtenus par l'instanciation de certaines variables d'une conjecture avec les ensembles couvrants de termes qu'on associe aux produit de leurs sortes sous la forme de *substitutions couvrantes*. Bachmair [3] avait montré que les ensembles couvrants de formules sont fondamentaux pour les preuves par cohérence. Ceci reste aussi vrai pour les preuves par récurrence implicite.

Afin de prouver des propriétés inductives, la procédure de récurrence par réécriture de termes n'a pas besoin d'être *réfutationnellement complète* pour garantir la détection en temps fini des conjectures fausses, comme c'est demandé dans [3], ni des systèmes de réécriture qui soient *confluents sur les termes clos* [3]. Kounalis et Rusinowitch [26] sont allés plus loin et ont proposé une technique de preuve par récurrence basée sur des *ensembles tests* [22] qu'on peut définir comme des ensembles couvrants de termes qui garantissent la réduction par réécriture des ensembles couvrants de formules générés.

Définition 6 (récurrence basée sur la réécriture de termes)**.** *Soit* $\{\varphi\sigma_1, \ldots, \varphi\sigma_n\}$ *les instances de la formule* $\forall \bar{x} \in \mathscr{E}, \varphi(\bar{x})$ *construites avec les substitutions couvrantes* $\sigma_1, \ldots, \sigma_n$*, respectivement. Le principe de la récurrence basé sur la réécriture de termes est défini par la formule :*

$$(\bigwedge_{i=1}^{n}(\bigwedge_{j=1}^{m_i} \varphi\theta_{m_i} <_f \varphi\sigma_i \Rightarrow \varphi\theta_{m_i}) \Rightarrow \varphi\sigma_i) \Rightarrow \forall \bar{x} \in \mathscr{E}, \varphi(\bar{x})$$

Comme les implémentations du principe de RNT, les implémentations du principe de récurrence basé sur la réécriture de termes ne peut pas gérer naturellement le raisonnement par récurrence mutuelle parce que les formules manipulées sont toutes des instances d'une seule formule. Cet inconvénient a été écarté par la proposition de la méthode de preuve par récurrence implicite, dont l'idée a été suggérée dans [26], puis présentée formellement dans [10].

3. Cependant, ces propriétés sont nécessaires pour *réfuter* les fausses conjectures.

Définition 7 (récurrence implicite)**.** *Soit $\mathscr{E}$ l'ensemble des conjectures produites par une procédure de récurrence implicite et qui termine par un ensemble vide de conjectures. Le principe de la récurrence implicite est une application immédiate du principe de RNF où la condition $(\forall m \in \mathscr{E}, (\forall k \in \mathscr{E}, k <_f m \Rightarrow k) \Rightarrow m)$ de (6) est garantie* implicitement *par la procédure de preuve.*

Pour une formule m arbitraire, si elle est trivialement vraie on n'a pas besoin d'hypothèses de récurrence pour sa preuve. Si par absurde m est fausse, alors on peut construire une suite strictement décroissante infinie de fausses formules de $\mathscr{E}$, ce qui contredit le fait que $<_f$ est nœthérien. Puisque la conjecture à prouver fait partie de $\mathscr{E}$, on conclut qu'elle est vraie.

La récurrence implicite peut être aussi mise en œuvre avec une généralisation des ensembles couvrants de formules appelés *ensembles couvrants contextuels* [33]. Le nombre de contraintes d'ordre à respecter peut également être réduit si on utilise un système de preuves *cyclique* [35], où on impose des contraintes d'ordre seulement aux conjectures qui, d'un point de vue de la théorie des modèles, dépendent mutuellement les unes des autres. Des travaux plus récents [36, 38, 40], laissent penser que la récurrence cyclique utilisée dans des cadres logiques de premier ordre autres que la logique équationnelle, comme celui qui inclut des prédicats inductifs [11], peut être expliquée par des principes de RNF. Quelquefois, les preuves par récurrence implicite sont plus automatiques que celles basées sur la récurrence explicite [6], dans d'autres cas il se produit le contraire [23]. D'autres analyses et comparaisons ont été faites dans [18, 21, 24, 30, 12, 43].

La RNF peut mieux gérer la récurrence mutuelle puisqu'elle accepte l'usage d'instances de différentes formules [34]. Elle permet aussi la récurrence *paresseuse* où les hypothèses de récurrence sont fournies sur demande et sont connues seulement au moment de leur application. En contrepartie, elle est peu utilisée par les systèmes de preuve actuels à cause de la difficulté de l'implémenter par une seule règle d'inférence et de gérer les dépendances d'ordre au niveau des preuves.

Relations entre les instances basées sur des termes et des formules

Nous allons maintenant étudier les rapports entre les deux types d'instances de premier ordre du principe de récurrence nœthérienne. Le théorème suivant est important aussi bien d'un point de vue théorique que pratique.

Théorème 2. *Tout principe de RNT peut être aussi représenté comme instance basée sur des formules.*

Démonstration. Soient $\mathscr{E}$ un ensemble non-vide de vecteurs de termes et φ une propriété vérifiée pout tout élément de $\mathscr{E}$ avec le principe de RNT suivant :

$$(\forall \overline{m} \in \mathscr{E}, (\forall \overline{k} \in \mathscr{E}, \overline{k} <_t \overline{m} \Rightarrow \varphi(\overline{k})) \Rightarrow \varphi(\overline{m})) \Rightarrow \forall \overline{p} \in \mathscr{E}, \varphi(\overline{p})$$

Soit $\mathscr{E}'$ l'ensemble $\{\varphi(\overline{p}) \mid \overline{p} \in \mathscr{E}\}$, on peut définir le principe de RNF comme suit :

$$(\forall \varphi(\overline{m}) \in \mathscr{E}', (\forall \varphi(\overline{k}) \in \mathscr{E}', \varphi(\overline{k}) <_f \varphi(\overline{m}) \Rightarrow \varphi(\overline{k})) \Rightarrow \varphi(\overline{m})) \Rightarrow \forall \varphi(\overline{p}) \in \mathscr{E}', \varphi(\overline{p})$$

où $\varphi(\overline{k}) <_f \varphi(\overline{m})$ si $\overline{k} <_t \overline{m}$. □

Corollaire 1. *Toute preuve par RNT peut être reconstruite en tant que preuve par RNF.*

D'un point de vue pratique, le raisonnement par RNT peut être *directement* intégré dans les systèmes de preuves qui implémentent le principe de RNF.

La traduction dans le sens inverse est aussi valable pour le cas suivant :

Théorème 3. *Le principe de RNF, définit par (6), peut être aussi représenté comme une instance basée sur des termes lorsque $\mathscr{E}$ est formé seulement par des instances d'une même formule.*

Démonstration. Soit le principe de RNF défini par la formule :

$$(\forall m \in \mathscr{E}, (\forall k \in \mathscr{E}, k <_f m \Rightarrow k) \Rightarrow m) \Rightarrow \forall p \in \mathscr{E}, p$$

où $\mathscr{E} = \{\varphi(\overline{x_1}), \ldots, \varphi(\overline{x_n}), \ldots\}$. Soit $\mathscr{E}'$ le nouvel ensemble $\{\overline{p} \mid \varphi(\overline{p}) \in \mathscr{E}\}$. On peut définir ainsi le principe de RNT :

$$(\forall \overline{m} \in \mathscr{E}', (\forall \overline{k} \in \mathscr{E}', \overline{k} <_t \overline{m} \Rightarrow \varphi(\overline{k})) \Rightarrow \varphi(\overline{m})) \Rightarrow \forall \overline{p} \in \mathscr{E}', \varphi(\overline{p})$$

où $\overline{k} <_t \overline{m}$ si $\varphi(\overline{k}) <_f \varphi(\overline{m})$. □

La validité du Théorème 3 pour le cas général reste un problème ouvert :

Conjecture 1. *Tout principe de RNF peut être aussi représenté en tant qu'instance basée sur des termes.*

Une conjecture similaire a été proposée dans [11], dans le cadre de la logique de premier ordre avec des définitions inductives, mais a été invalidée ultérieurement [4].

Certification avec Coq du raisonnement par RNF

La formalisation en Coq du principe de RNF suit des idées présentées dans [34, 41] et utilisées pour certifier automatiquement des preuves par récurrence implicite développées avec le démonstrateur SPIKE [39]. D'abord, on associe à chaque formule une *mesure* qui va nous permettre de comparer des formules et de définir l'ordre de récurrence. Étant donné une liste LF de paires de la forme $(\varphi,\ \mu_\varphi)$, où φ est une formule et μ_φ sa mesure, (6) peut être reformulée comme suit :

$$(\forall p \in \mathsf{LF}, (\forall p' \in \mathsf{LF}, snd(p') <_f snd(p) \Rightarrow fst(p')) \Rightarrow fst(p)) \Rightarrow \forall f \in \mathscr{E}, fst(f),$$

où *fst* (resp., *snd*) est la fonction qui retourne la première (resp., deuxième) projection d'une paire. Dans Coq, la partie conditionnelle de l'implication la plus extérieure peut être formalisée par le lemme suivant :

`Lemma` main : $\forall p$, In p LF $\rightarrow$ ($\forall p'$, In p' LF $\rightarrow$ less (snd p') (snd p)) $\rightarrow$ fst p') $\rightarrow$ fst p.

L'ordre de récurrence, noté less, est une implémentation de $<_f$ en tant qu'extension multi-ensemble d'un ordre rpo [2] sur les termes. La mesure d'une formule peut être représentée par un multi-ensemble de ses sous-termes. Étant donné deux paires $(\varphi_1, \mu_{\varphi_1})$ et $(\varphi_2, \mu_{\varphi_2})$, on dit que φ_1 est *plus petit que* φ_2 si less μ_{φ_1} μ_{φ_2}.

Les étapes de récurrence principales pour prouver la validité d'une formule φ de chaque paire de LF sont : i) la partie déductive : choisir des (instances de) formules de LF en tant qu'hypothèses de récurrence à utiliser dans la preuve de φ, et ii) la partie des contraintes d'ordre : montrer que ces hypothèses de récurrence sont plus petites que φ. La partie ii) peut être omise si la partie i) ne nécessite pas de raisonnement par récurrence.

L'ordre less a été mis en œuvre avec les représentations syntaxiques des termes fournies par la bibliothèque COCCINELLE [14, 15] qui permet de modéliser des notions mathématiques pour la réécriture, comme les algèbres de termes et les ordres de récurrence, ainsi que par la bibliothèque CoLoR [5], utilisée pour implémenter la relation « extension multi-ensemble ».

Les mesures doivent suivre les instanciations des formules. C'est pour cette raison que la liste LF du lemme main a été adaptée pour inclure des fonctions anonymes à la place des paires afin de pouvoir partager les variables communes entre les formules et leurs mesures :

`Definition` type_LF := *argument_sort* $\rightarrow$ (`Prop` $\times$ (**List.list term**)),

où *argument_sort* est la sorte écrite sous forme curryfiée de la plus générale version du vecteur de variables partagées par chaque paire de LF. La définition du lemme main dans la nouvelle mouture est :

`Lemma` main : $\forall f$, In f LF $\rightarrow$ $\forall u$, ($\forall f'$, In f' LF $\rightarrow$ $\forall u'$, less (snd (f' u')) (snd (f u)) $\rightarrow$ fst (f' u')) $\rightarrow$ fst (f u).

Toute formule de la liste LF, dont les conjectures initiales, peut être certifiée si on prouve le théorème all_true. Ceci nécessite l'utilisation du lemme main et du principe de récurrence nœthérienne (1) qui est intégré nativement dans Coq :

`Theorem` all_true : $\forall f$, In f LF $\rightarrow$ $\forall u$: **nat**, fst (f u).

Conclusions et travaux à venir

Nous avons présenté une taxonomie du raisonnement par récurrence nœthérienne de premier ordre, ainsi qu'un survol de ses instances les plus emblématiques : basées sur des termes et des formules. La traduction des preuves intégrant des instances basées sur des formules $\rightarrow$ termes reste un problème ouvert. On a montré sa validité seulement pour certains cas (voir Théorème 3). Nous sommes en train d'étudier d'autres cas, par exemple, les preuves avec des séquents dans des logiques multi-sortées de premier ordre avec des prédicats inductifs. Nous n'avons pas discuté des problématiques importantes de la preuve par récurrence, comme la réfutation des fausses conjectures, le choix des variables de récurrence, l'usage des hypothèses de récurrence, l'ajout de nouveaux lemmes, la généralisation des conjectures, ainsi que le choix des stratégies de preuve. Le lecteur intéressé trouvera plus de détails ailleurs [19].

La formalisation en Coq du principe de RNF nous a permis de certifier des preuves par récurrence implicite générées automatiquement par SPIKE, concernant entre autres des lemmes cruciaux pour valider la conformité d'un protocole de télécommunications [32]. On envisage d'appliquer la même méthodologie de certification sur des preuves cycliques, comme celles produites avec E-CYCLIST [40].

Références

[1] R. AUBIN : Mechanizing structural induction. *Theor. Comput. Sci.*, 9:329–362, 1979.

[2] F. BAADER et T. NIPKOW : *Term Rewriting and All That*. Cambridge University Press, 1998.

[3] L. BACHMAIR : Proof by consistency in equational theories. *Logic in Computer Science, 1988. LICS '88., Proceedings of the Third Annual Symposium on*, pages 228–233, 1988.

[4] S. BERARDI et M. TATSUTA : Classical system of Martin-Lof's inductive definitions is not equivalent to cyclic proofs. *Logical Methods in Computer Science*, 15(3), 2019.

[5] F. BLANQUI et A. KOPROWSKI : CoLoR : a Coq library on well-founded rewrite relations and its application to the automated verification of termination certificates. *MSCS*, 21(4):827–859, 2011.

[6] A. BOUHOULA et M. RUSINOWITCH : Implicit induction in conditional theories. *Journal of Automated Reasoning*, 14(2):189–235, 1995.

[7] R. BOULTON et K. SLIND : Automatic derivation and application of induction schemes for mutually recursive functions. *In* J. LLOYD, V. DAHL, U. FURBACH, M. KERBER, K.-K. LAU, C. PALAMIDESSI, L. PEREIRA, Y. SAGIV et P. STUCKEY, éditeurs : *Computational Logic — CL 2000*, volume 1861 de *Lecture Notes in Computer Science*, pages 629–643. Springer Berlin / Heidelberg, 2000.

[8] R. S. BOYER et J S. MOORE : *A Computational Logic*. Academic Press, New York, NY, 1979.

[9] R. S. BOYER et J S. MOORE : *A Computational Logic Handbook*. Academic Press Professional, 1988.

[10] F. BRONSARD, U.S. REDDY et R. HASKER : Induction using term orderings. *In CADE (Conf. on Automated Deduction)*, volume 814 de *LNCS*, pages 102–117. Springer, 1994.

[11] J. BROTHERSTON et A. SIMPSON : Sequent calculi for induction and infinite descent. *Journal of Logic and Computation*, 21(6):1177–1216, 2011.

[12] H. COMON : Inductionless induction. *In* A. ROBINSON et A. VORONKOV, éditeurs : *Handbook of Automated Reasoning*, pages 913–962. Elsevier and MIT Press, 2001.

[13] H. COMON et R. NIEUWENHUIS : Induction= I-axiomatization+ first-order consistency. *Information and Computation(Print)*, 159(1-2):151–186, 2000.

[14] E. CONTEJEAN, P. COURTIEU, J. FOREST, O. PONS et X. URBAIN : Certification of automated termination proofs. *Frontiers of Combining Systems*, pages 148–162, 2007.

[15] E. CONTEJEAN, A. PASKEVICH, X. URBAIN, P. COURTIEU, O. PONS et J. FOREST : A3PAT, an approach for certified automated termination proofs. *In* J. P. GALLAGHER et J. VOIGTLÄNDER, éditeurs : *PEPM - Proceedings of the 2010 ACM SIGPLAN Workshop on Partial Evaluation and Program Manipulation, PEPM 2010, Madrid, Spain*, pages 63–72. ACM, 2010.

[16] N. DERSHOWITZ : Applications of the Knuth-Bendix completion procedure. *In Seminaire d'Informatique Theorique*, pages 95–111, 1982.

[17] L. FRIBOURG : A strong restriction of the inductive completion procedure. *Journal of Symbolic Computation*, 8(3):253 – 276, 1989.

[18] S. J. GARLAND et J. V. GUTTAG : Inductive methods for reasoning about abstract data types. *Proceedings of the 15th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 219–228, 1988.

[19] B. GRAMLICH : Strategic issues, problems and challenges in inductive theorem proving. *Electronic Notes in Theoretical Computer Science*, 125(2):5–43, 2005.

[20] G. HUET et J.M. HULLOT : Proofs by induction in equational theories with constructors. Rapport technique 0028, INRIA, 1980.

[21] J.-P. JOUANNAUD et E. KOUNALIS : Automatic proofs by induction in theories without constructors. *Information and Computation*, 82(1):1 – 33, 1989.

[22] D. KAPUR, P. NARENDRAN et H. ZHANG : Proof by induction using test sets. *In 8th International Conference on Automated Deduction*, volume 230 de *Lecture Notes Computer Science*, pages 99–117. Springer, 1986.

[23] D. KAPUR et M. SUBRAMANIAM : Automating induction over mutually recursive functions. *In Algebraic Methodology and Software Technology*, volume 1101 de *LNCS*, pages 117–131. Springer, 1996.

[24] D. KAPUR et H. ZHANG : Automating induction : Explicit vs. inductionless. *Proc. Third International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, Florida*, pages 2–5, 1994.

[25] D.E. KNUTH et P.B. BENDIX : Simple word problems in universal algebras. *In Computational Problems in Abstract Algebra*, pages 263–297, 1970.

[26] E. KOUNALIS et M. RUSINOWITCH : Mechanizing inductive reasoning. *In Proceedings of the Eighth National Conference on Artificial Intelligence - Volume 1*, AAAI'90, pages 240–245. AAAI Press, 1990.

[27] W. KÜCHLIN : Inductive completion by ground proof transformation. *In* H. AIT-KACI et M. NIVAT, éditeurs : *Resolution of Equations in Algebraic Structures (Volume II) : Rewriting Techniques*, pages 211–244. Academic Press, London, 1989.

[28] J. MCCARTHY et J. PAINTER : Correctness of a compiler for arithmetic expressions. *In Mathematical Aspects of Computer Science*, pages 33–41. American Mathematical Society, 1967.

[29] D. R. MUSSER : On proving inductive properties of abstract data types. *In POPL*, pages 154–162, 1980.

[30] D. NAIDICH : On generic representation of implicit induction procedures. Rapport technique CS-R9620, CWI, 1996.

[31] U.S. REDDY : Term rewriting induction. *Proceedings of the 10th International Conference on Automated Deduction*, pages 162–177, 1990.

[32] M. RUSINOWITCH, S. STRATULAT et F. KLAY : Mechanical verification of an ideal incremental ABR conformance algorithm. *Journal of Automated Reasoning*, 30(2):153–177, 2003.

[33] S. STRATULAT : A general framework to build contextual cover set induction provers. *Journal of Symbolic Computation*, 32(4):403–445, 2001.

[34] S. STRATULAT : Integrating implicit induction proofs into certified proof environments. *In IFM'2010 (8th International Conference on Integrated Formal Methods)*, volume 6396 de *Lecture Notes in Computer Science*, pages 320–335, 2010.

[35] S. STRATULAT : A unified view of induction reasoning for first-order logic. *In* A. VORONKOV, éditeur : *Turing-100 (The Alan Turing Centenary Conference)*, volume 10 de *EPiC Series*, pages 326–352. EasyChair, 2012.

[36] S. STRATULAT : Cyclic proofs with ordering constraints. *In* R. A. SCHMIDT et C. NALON, éditeurs : *TABLEAUX 2017 (26th International Conference on Automated Reasoning with Analytic Tableaux and Related Methods)*, volume 10501 de *LNAI*, pages 311–327. Springer, 2017.

[37] S. STRATULAT : Mechanically certifying formula-based Noetherian induction reasoning. *Journal of Symbolic Computation*, 80, Part 1:209–249, 2017.

[38] S. STRATULAT : Validating back-links of $\mathrm{FOL}_{\mathrm{ID}}$ cyclic pre-proofs. *In* S. BERARDI et S. van BAKEL, éditeurs : *CL&C'18 (Seventh International Workshop on Classical Logic and Computation)*, numéro 281 de EPTCS, pages 39–53, 2018.

[39] S. STRATULAT : SPIKE, an automatic theorem prover – revisited. *In SYNASC 2020 : Proceedings of the 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, pages 93–96. IEEE Computer Society, 2020.

[40] S. STRATULAT : E-Cyclist : Implementation of an efficient validation of $\mathrm{FOL}_{\mathrm{ID}}$ cyclic induction reasoning. *In* T. KUTSIA, éditeur : *9th International Symposium on Symbolic Computation in Software Science*, volume 342 de *Electronic Proceedings in Theoretical Computer Science*, pages 129–135, juin 2021.

[41] S. STRATULAT et V. DEMANGE : Automated certification of implicit induction proofs. *In CPP'2011 (First International Conference on Certified Programs and Proofs)*, volume 7086 de *Lecture Notes Computer Science*, pages 37–53. Springer Verlag, 2011.

[42] THE COQ DEVELOPMENT TEAM : *The Coq Reference Manual*. INRIA, 2020. `http://coq.inria.fr/doc`.

[43] C.-P. WIRTH : History and future of implicit and inductionless induction : Beware the old jade and the zombie ! *In Mechanizing Mathematical Reasoning : Essays in Honor of Jörg H. Siekmann on the Occasion of His 60th Birthday*, numéro 2605 de Lecture Notes in Artificial Intelligence, pages 192–203. Springer, 2005.

[44] H. ZHANG, D. KAPUR et M. S. KRISHNAMOORTHY : A mechanizable induction principle for equational specifications. *In Proceedings of the 9th International Conference on Automated Deduction*, pages 162–181, London, UK, 1988. Springer-Verlag.

Étonnantes puissances de 2

Jean-Paul Delahaye [1]

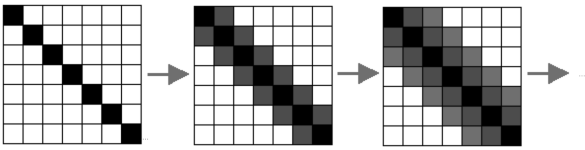
La rubrique « Récréation informatique » propose une petite énigme algorithmique ou sur un thème de mathématiques discrètes susceptible d'intéresser un lecteur de 1024. La solution est donnée dans le numéro suivant.

Rappel et solution du problème précédent

L'INFECTION DU DAMIER

Un damier de taille carrée et de n cases de côté subit une infection dont la règle de fonctionnement est la suivante : si une case non infectée a au moins deux voisines infectées, elle l'est à la seconde suivante. Ne comptent comme voisines d'une case donnée que la case en dessous, la case au-dessus, la case à gauche, et la case à droite.

Si on suppose qu'à un instant donné toutes les cases d'une diagonale sont infectées alors on comprend que progressivement toutes les cases se trouveront infectées.



1. Professeur émérite, université de Lille, campus scientifique, CRIStAL UMR CNRS, 9189 Centre de recherche en informatique signal et automatique de Lille, bâtiment ESPRIT, 59655, Villeneuve d'Ascq Cedex France. E-mail : jean-paul.delahaye@univ-lille.fr.

DOI:10.48556/SIF.1024.19.171

Le mécanisme d'infection semble particulièrement violent et rapide, pourtant tout n'est pas possible, et même si cela semble paradoxal, pour infecter le damier de côté n dans son entier, il faut, qu'au départ, il y ait au moins n cellules infectées.

Pouvez-vous le démontrer ? Le plus étonnant dans ce problème est qu'un seul mot donne la démonstration recherchée.

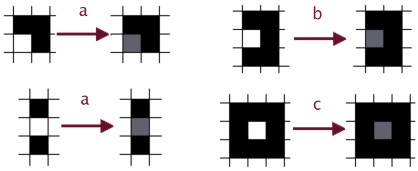
SOLUTION.

Merci à Marie Davenne et Jean-Michel Batto qui m'ont fait parvenir les bonnes solutions. Le mot qui donne la solution la plus simple est « périmètre ». En effet, quand une case est infectée, trois cas sont possibles :

a) le périmètre de l'ensemble des cases infectées augmente de deux unités et diminue de deux unités, donc reste stable ; cette situation se présente lorsque la nouvelle case infectée est entourée exactement de deux cases déjà infectées ;

b) le périmètre diminue de 2 ; cette situation se présente lorsque la nouvelle case infectée est entourée de trois cases infectées ;

c) le périmètre diminue de 4 ; cette situation se présente lorsque la nouvelle case infectée est entourée par ses quatre côtés.



Le périmètre de l'ensemble des cases infectées n'augmente donc jamais. Le périmètre du damier est $4n$, il faut donc, pour infecter tout le damier, avoir au moins n cases infectées au départ et qu'elles ne se touchent pas.

Nouveau problème

ÉTONNANTES PUISSANCES DE 2

Comme tout informaticien, vous êtes fascinés par les puissances de 2 : 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048... Si on observe le dernier chiffre (chiffre des unités), il est immédiat de constater et de démontrer que la série '2', '4', '8', '6' y est répétée indéfiniment : le dernier chiffre est donc dans 25 % des cas un '2', dans 25 % des cas un '4', dans 25 % des cas un '8' et dans 25 % des cas un '6'. Si on observe les deux derniers chiffres, on remarque (et on démontre sans mal encore) que la séquence '04', '08', '16', '32', '64', '28', '56', '12', '24', '48', '96', '92', '84', '68', '36', '72', '44', '88', '76', '52' revient périodiquement. On en déduit que le deuxième chiffre à partir de la droite (chiffre des dizaines) est périodiquement '0',

’0’, ’1’, ’3’, ’6’, ’2’, ’5’, ’1’, ’2’, ’4’, ’9’, ’9’, ’8’, ’6’, ’3’, ’7’, ’4’, ’8’, ’7’, ’5’. Cette séquence périodique de longueur 20 comporte chaque chiffre deux fois exactement, ce qui a pour conséquence que le ’0’ sera le deuxième chiffre de 2^k à partir de le droite (chiffre des dizaines) dans 10 % des cas, et que c’est vrai aussi pour le ’1’, le ’2’, etc. Les chiffres en deuxième position à partir de la droite dans 2^k sont présents chacun avec une fréquence de 10 %. L’étude du chiffre en troisième position à partir de la droite (chiffre des centaines) conduit à la même conclusion, et il en va de même aussi pour le quatrième, etc. La propriété n’est pas facile à démontrer mais elle est vraie et avec un programme, on la vérifie expérimentalement [1].

Il est donc vrai que sauf pour le chiffre des unités, les chiffres des puissances de 2 placés en position n à partir de la droite ont chacun une fréquence d’apparition de 10 % et cela, quel que soit $n > 1$.

Considérons maintenant le second chiffre à gauche des 2^n qui peut être 0, 1, ..., 9. Une étude numérique montre, cette fois, que les dix chiffres n’ont pas la même fréquence d’apparition :

0	1	2	3	4	5	6	7	8	9
0.120	0.114	0.109	0.104	0.100	0.097	0.093	0.090	0.087	0.084

FIGURE 1. Fréquence d’apparition du second chiffre à gauche de 2^n.

Parfois c’est plus que 0.1, parfois moins. Nous n’avons pas considéré le premier chiffre à gauche car il ne peut pas être ’0’ et donc, nous ne pouvions pas espérer avoir une fréquence d’apparition égale à 0.1.

N’est-il pas paradoxal que l’égalité des fréquences des 10 chiffres qu’on constate et qu’on démontre pour le chiffre des dizaines, pour le chiffres des centaines, etc. n’entraîne pas l’égalité des fréquences pour ce second chiffre à gauche ?

Comment l’expliquer ?

Envoyez vos réponses à jean-paul.delahaye@univ-lille.fr. Les noms des premiers lecteurs à me donner la bonne réponse (et à la justifier) seront mentionnés dans le prochain numéro de 1024.

Références

[1] Michael Fuchs Digital Expansion of Exponential Sequences J. Theor. Nombres Bordeaux, 14 :2, 477-487, 2002.

ADHÉRER À LA SIF

La Société informatique de France est un espace de réflexion, de concertation sur les enjeux de l'informatique, mais aussi un espace d'actions, basé sur le travail de la communauté, qui vise à rassembler toutes celles et tous ceux pour qui faire progresser l'informatique est un métier ou une passion : enseignants, chercheurs, ingénieurs, industriels, consultants et étudiants.

Les adhésions sont valables 12 mois à compter de la date d'adhésion.

Personnes physiques

Tarif plein : 30 €
Tarif réduit : 15 €

- Membre d'un adhérent institutionnel de la SIF.
- CDD, CDI depuis moins de 2 ans, retraité.
- Membre d'une association partenaire, ou de l'ACM.

Gratuit : étudiants, doctorants et post-doctorants.
La SIF vous offre la possibilité d'effectuer le règlement de la cotisation directement en ligne.

Partenaires (Personnes morales)

Les associations partenaires, membres du Conseil des associations de la SIF, ne paient pas de cotisation. Les institutions telles que laboratoires, unités d'enseignement, ou entreprises, peuvent adhérer en tant que telles à la SIF. Il n'existe pas de tarif spécifique pour les adhérents institutionnels : en fonction de leur taille, de leur secteur d'activité, l'importance de l'effort ne se mesure pas de la même façon.

La SIF propose cinq niveaux de cotisation. Pour vous aider dans votre choix, vous trouverez ci-dessous une indication du tarif en fonction de la taille de l'adhérent institutionnel :

- Tarif 1 : 250 € (moins de 50 personnes)
- Tarif 2 : 500 € (de 50 à 100 personnes)
- Tarif 3 : 1000 € (de 100 à 150 personnes)
- Tarif 4 : 1500 € (de 150 à 200 personnes)
- Tarif 5 : 2000 € (au delà de 200 personnes)

Pour adhérer, vous devez contacter notre trésorier. Pour toute question, ne pas hésiter à contacter notre secrétariat.

Plus d'informations sur notre site internet :

`https://www.societe-informatique-de-france.fr`

de la société informatique
de France

Institut Henri Poincaré,
11 rue Pierre et Marie Curie,
75231 Paris Cedex