

Déploiement du cahier de laboratoire électronique à l'INSERM et nouvelles perspectives

Paul Guy Dupré

Inserm DSI
13 rue Watt
75013 Paris

Claudia Gallina-Muller

Inserm DSI
9 avenue de la Forêt de Haye
54505 Vandœuvre-lès-Nancy

Résumé

Le DSI de l'Inserm s'intéresse, depuis 2013, aux solutions de cahiers de laboratoire électroniques (ELN, Electronic Laboratory Notebook). La phase d'expérimentation a été présentée aux JRES 2017. Nous avons alors proposé une définition de l'ELN et expliqué les enjeux. La solution sert à décrire et documenter toute la phase confidentielle de réalisation d'un projet de recherche. Certains enjeux perdurent, le cahier de laboratoire doit répondre aux obligations légales et contractuelles, notamment en apportant la preuve de l'invention et de ses inventeurs. Les apports du numérique sont multiples. L'ELN améliore la traçabilité des recherches, la lutte contre la fraude et la gestion des données. Il facilite les démarches qualité et le dépôt des brevets. Il peut apporter de nouvelles fonctionnalités, telles que le travail collaboratif, l'accès distant, la gestion des projets, la gestion documentaire et le contrôle d'accès.

CLÉ, le Cahier de Laboratoire Électronique de l'Inserm, propose depuis 2018 une solution dédiée à la biologie, prenant également en charge la gestion des inventaires et des équipements. Nous détaillerons les aspects appels d'offres, architecture, et conduite du changement de ce projet.

L'ELN se retrouve en position centrale dans le SI du laboratoire. Il peut apporter des solutions pour répondre aux enjeux du moment, améliorer la gestion de la donnée et la reproductibilité des expériences. CLÉ peut également s'interfacer avec des bases de données du laboratoire et à notre catalogue de service (bibliographie, gestion financière, feuille de temps, gestion des risques, agendas, stockage des données brutes, GED).

En conclusion, nous évoquerons les possibilités d'interfaçage avec des outils de traitement des données de recherche.

Mots-clefs

Cahier de laboratoire électronique, ELN, LIMS, Démarche qualité, Gestion des données, DMP

1. Introduction

La plateforme d'expérimentation de l'ELN (ELN, Electronic Laboratory Notebook), mise en place à l'Inserm entre 2015 et 2018, a permis de tester une solution avec une trentaine de laboratoires. Ces aspects avaient été exposés lors d'une communication aux JRES 2017 [1]. Cette longue phase d'expérimentation nous a permis de recueillir les avis, les besoins d'évolutions, d'évaluer dans la durée les apports mais aussi les difficultés de mise en œuvre dans les laboratoires pilotes. Cette période nous a permis d'acquérir une expertise sur ce type de solutions, de valider des choix et de mettre en place un plan de déploiement.

CLÉ est la solution Inserm de Cahier de Laboratoire Électronique, de gestion des stocks, des collections et des instruments associés. Elle est utilisée pour détailler au quotidien l'ensemble des travaux conduits par les personnels de la recherche. Ses apports vont bien au-delà d'une simple démarche de dématérialisation. Le numérique facilite la traçabilité complète des expériences (référence aux bases d'inventaires et d'instruments), la gestion et la préservation de la connaissance. L'ELN apporte des fonctionnalités supplémentaires, telles que l'accès distant, le travail collaboratif, le management des projets.

Ce retour d'expérience a été largement partagé à l'Inserm. Nous avons pu associer tous les acteurs au cadrage du projet, à la définition du périmètre, aux propositions stratégiques. L'étape finale de cette expérimentation a été la validation, par la Direction Générale, des choix pour l'établissement.

Dans le cadre du projet CLÉ, la solution Labguru de l'éditeur Biodata est proposée depuis décembre 2018 à l'ensemble des laboratoires Inserm. Nous reviendrons dans cet article sur toutes les phases du projet (étude, réalisation, déploiement). Nous aborderons ensuite les opportunités pour que les ELN répondent à de nouveaux enjeux.

2. CLÉ, le projet Inserm de Cahier de Laboratoire Électronique

2.1. Planning

Les phases d'étude et de réalisation du projet se sont déroulées sur deux ans.

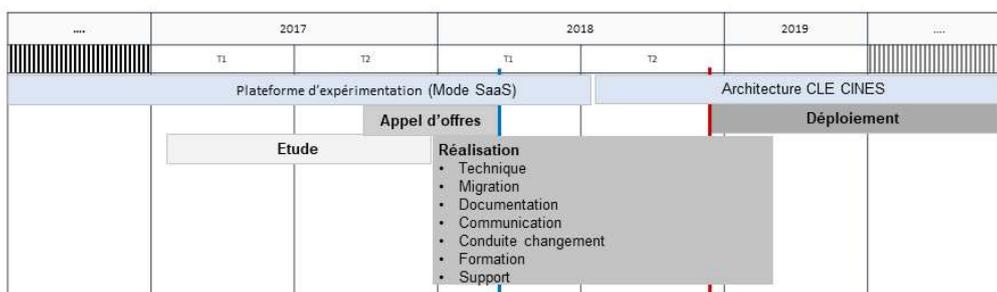


Figure 1- CLÉ: planification de réalisation

2.2. Étude

2.2.1. Besoins exprimés par les Laboratoires

Les demandes récurrentes de cahier de laboratoire électronique ont émergé à l'Inserm à partir de 2013. Les premiers laboratoires demandeurs étaient le plus souvent engagés dans une démarche qualité ou soumis à des exigences spécifiques de partenaires industriels. Les premiers enjeux exprimés étaient la traçabilité, la reproductibilité des expériences et la conservation.

Dès l'origine, l'étude a été réalisée avec des laboratoires pilotes issus de l'ensemble des thématiques de recherche portées par l'Inserm. C'est la notion même de laboratoire qui a fait l'objet d'une attention particulière pour définir le périmètre d'un cahier, la notion de groupe qui lui est associée, son responsable, ses collaborations et la confidentialité des données au sein du groupe. Un deuxième sujet majeur a été l'intégration des outils existants au cahier (intégration des données en général et des fichiers Microsoft Office en particulier).

2.2.2. Apports pour l'établissement

Nous avons plus largement travaillé avec l'ensemble des acteurs pour définir les apports attendus pour l'Inserm :

- répondre aux obligations légales ou contractuelles ;
- faciliter les procédures de dépôt de brevets ;
- renforcer le sentiment d'appartenance ;
- faciliter les démarches qualité et proposer un outil améliorant la traçabilité ;
- disposer d'un outil de gestion de la connaissance et des données biologiques ;
- sécuriser les données scientifiques ;
- lutter contre la fraude en assurant la traçabilité rétrospective et l'horodatage ;
- permettre le travail collaboratif et les accès distants ;
- faciliter le management des équipes et des projets ;
- uniformiser au sein du laboratoire la description des projets, des expériences et des protocoles.

2.2.3. Analyse de sécurité

Un des principaux risques détectés au cours de l'analyse concerne la sensibilité des données des expériences et celle des protocoles. Il nous est apparu important de pouvoir gérer finement les aspects de confidentialité au niveau du laboratoire et de confier à son responsable opérationnel PI (Principal Investigator), la gestion des droits d'accès et d'export des données.

2.2.4. Analyse juridique

Cette nouvelle offre proposée par l'Inserm ne modifie en rien la propriété du contenu et sa valorisation, qui sont définies par l'ensemble des partenaires dans le cadre de la convention de mixité des UMR (Unité Mixte de Recherche).

Cette offre est opérée par l’Inserm en conformité avec la législation et notamment le RGPD¹ (Règlement Général sur la Protection des Données). Le traitement des données personnelles, mis en œuvre dans le cadre du cahier de laboratoire électronique, est nécessaire au respect des obligations légales de l’Inserm. Cette conformité a été étudiée et validée par la DPO (Data Protection Officer) de l’Inserm. Elle concerne les éléments contractuels avec nos prestataires, les processus mis en œuvre pour assurer la protection des données personnelles, ainsi que la notice d’information des utilisateurs.

La conservation des cahiers de laboratoire est organisée à la fois pour des raisons juridiques et des raisons patrimoniales :

- pour garantir le respect de ses obligations légales, l’Inserm conservera par défaut l’ensemble des cahiers de laboratoire pendant une durée de 25 ans. C’est la période la plus longue qui correspond au cas d’un brevet, sachant que celui des contrats industriels est classiquement de 10 ans après la date de la dernière publication ;
- les données de recherche de l’Inserm sont des archives publiques indestructibles, imprescriptibles et inaliénables (Code du patrimoine, art. L. 211-1.). La conservation de ces archives sur du long terme sera organisée par le service des archives de l’Inserm.

Les utilisateurs doivent se conformer à la réglementation et aux conditions générales d'utilisation du service. Une vigilance particulière est apportée aux types de données pouvant être hébergées dans le cahier de laboratoire et notamment les données de santé qui ne sont généralement pas éligibles à notre solution, qui n’a pas les agréments nécessaires.

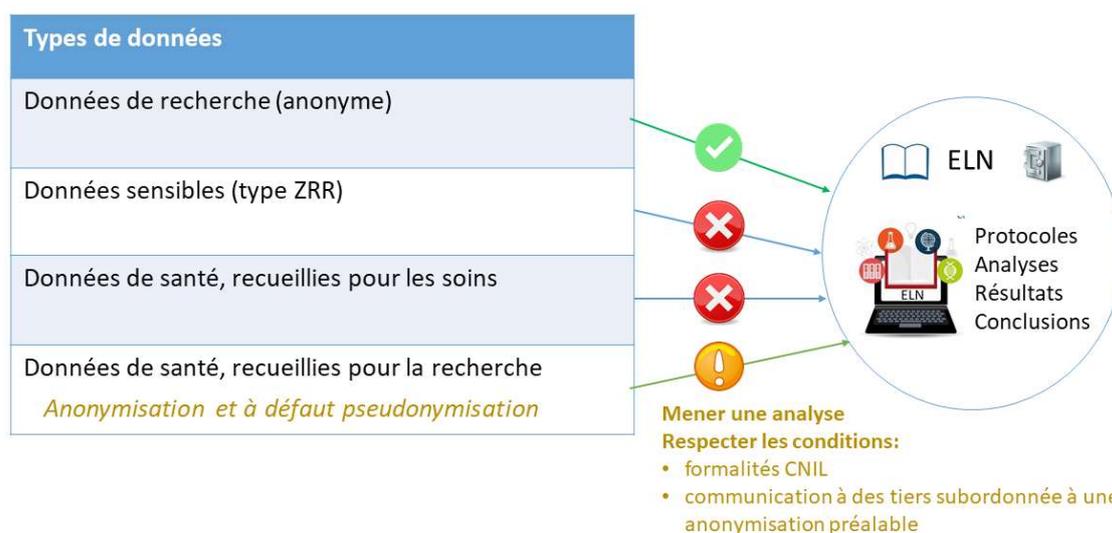


Figure 2 - Éligibilité des données de santé

2.2.5. Choix stratégiques

Pour garantir son autonomie et la sécurité de ses données, l’Inserm a souhaité héberger la solution, maîtriser l’accès aux données et mettre en place la réversibilité.

Pour assurer la pérennité de la solution et garantir la stabilité des tarifs, nous avons réalisé un marché sur une durée de 4 ans.

Pour garantir la protection et le contrôle sur les données, nous avons imposé une organisation par laboratoire. Son responsable devra garantir un strict contrôle de la confidentialité des données (autoriser

¹ <https://www.cnil.fr/fr/reglement-europeen-protection-donnees>

l'accès et fixer les règles de confidentialité des données et d'export). Nous avons également défini les mesures à mettre en œuvre au niveau central.

Voici les principales fonctionnalités exprimées:

- collaboratives ;
- proposant des outils dédiés à la biologie, à la chimie et une gestion des inventaires ;
- ouvertes sur le SI du laboratoire (Import / Export, compatibilité Office, API) ;
- organisées autour d'un responsable de laboratoire PI qui administre localement la solution ;
- supports et formations bilingues ;
- authentification shibbolethisée .

2.3. Appel d'offres

Le marché mis en place pour une durée de 4 ans concernait une tranche annuelle ferme comprenant les licences, l'installation, des aspects d'exploitation et de maintenance, ainsi que l'assistance et la formation des administrateurs et utilisateurs de la solution.

Pour rendre notre appel d'offres attractif et ouvert à l'international, l'Inserm a publié en français et en anglais et a laissé la possibilité aux prestataires de répondre dans ces deux langues. Le CCTP était précis, concis, avec un minimum de termes équivoques. Nous avons focalisé notre demande uniquement sur nos objectifs et les fonctionnalités importantes souhaitées en nous engageant financièrement pour une durée de 3 ans minimum. L'appel d'offres a été publié en décembre 2017. Seules deux réponses conformes à la procédure ont été analysées.

Sur le critère de la valeur technique de l'offre, noté à 60% du total, l'offre de Biodata répondait légèrement mieux aux besoins exprimés, principalement sur les aspects de collaboration dans le laboratoire. Sur le critère du prix, noté à 40%, Biodata était également mieux disant.

2.4. Phase d'initialisation

2.4.1. Architecture

La mise en service de l'application CLÉ sur les infrastructures de l'Inserm a nécessité la mise en place de 6 machines virtuelles (VM) sous Linux. L'architecture de CLÉ a été définie pour 13000 utilisateurs. Afin d'optimiser les performances, les serveurs sont équipés de disques SSD.

Quatre VM hébergent les briques applicatives dans différents conteneurs Docker. Deux VM hébergent la base de données principale sous MySQL (VM principale : 24 cœurs, 32 Go de RAM et 2 TB SSD).

L'ensemble des flux passe par un proxy. L'authentification est effectuée via la fédération d'identité de Renater (REseau NATIONAL de Télécommunications pour la technologie, l'Enseignement et la Recherche). Un Service Provider (SP) a été intégré à l'application.

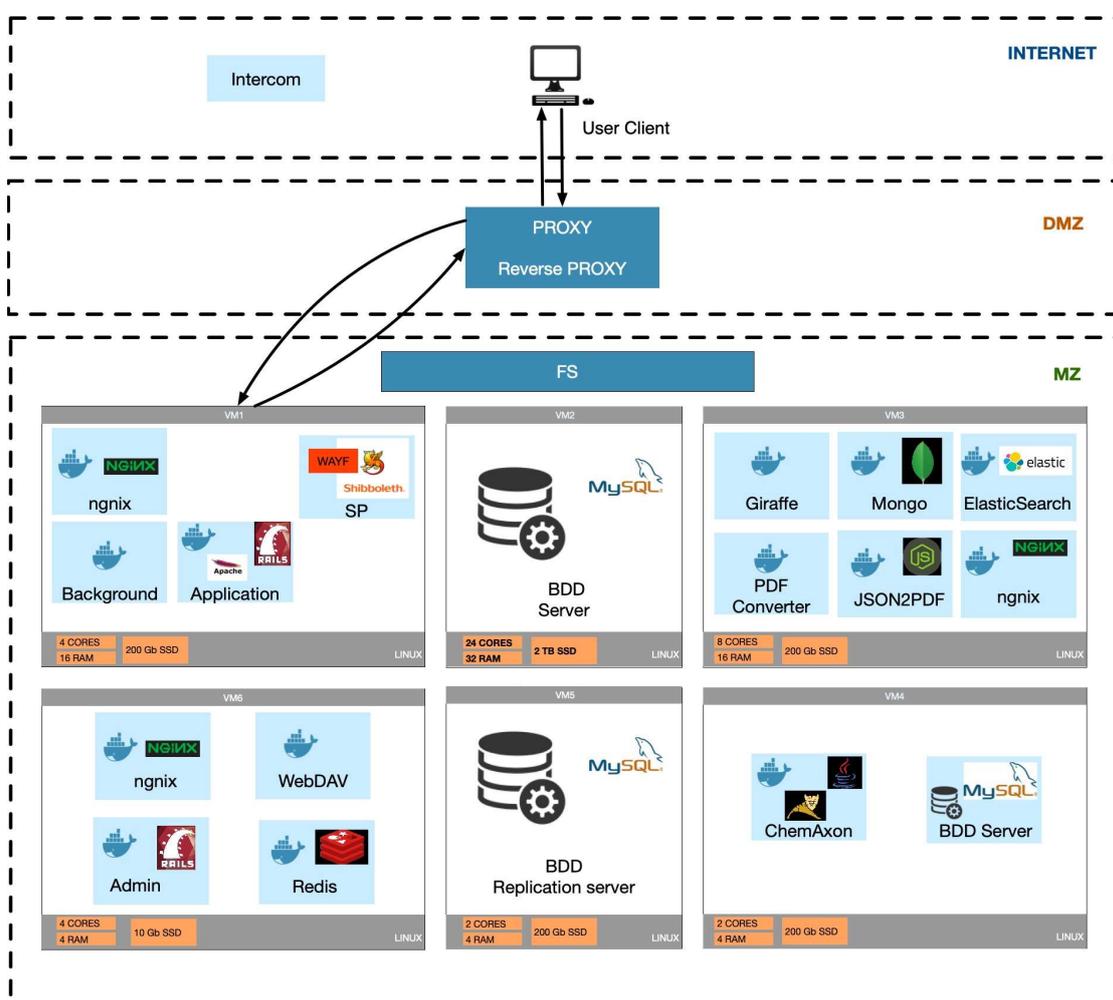


Figure 3 - Architecture et composants de CLÉ

	Ruby on Rails, framework web.		Elasticsearch, serveur d'indexation et de recherche de données.
	Apache, serveur HTTP.		Plateforme JavaScript Node.js pour le développement d'API.
	MySQL, base de donnée SQL.	 	Apache Tomcat, conteneur web libre de servlets et JSP. Java JSP(Java Server Pages).
	MongoDB, base de données orientée documents (NoSQL).		Nginx, serveur HTTP, cache, load balancer
	Redis, base de données clef-valeur scalable, hautes performances, et distribuée (NoSQL).		

Figure 4 - Composants applicatifs

2.4.2. Installation

La phase d'installation a nécessité la coordination des équipes internes et du prestataire pour la mise en place préalable des 6 VM, des conteneurs Docker, des bases MySQL, ainsi que pour l'identification et l'ouverture des flux entre les différents composants. La base de données principale a été installée sur deux VM. Le reste des composants est fourni sous forme d'images Docker. Les vérifications de conformité ont ensuite été réalisées et des procédures de sauvegarde ont été mises en place.

Une fois l'installation terminée, les données des laboratoires pilotes présentes sur la plateforme d'expérimentation externalisée ont été récupérées en production.

La shibbolethisation de l'application a été une étape difficile à mettre en œuvre. L'éditeur a ajouté le composant Service Provider (SP). La compréhension pour tous du fonctionnement de la fédération d'identité de Renater a été une longue étape. Au-delà de cette étape de paramétrage, le bon fonctionnement de l'authentification reste dépendant de chaque IDP (Identity Provider) d'établissement. Ceci complique le support aux utilisateurs. Dans le cas de CLÉ, nous utilisons le champ «email» lors des étapes de validation de la connexion. Tous les IDP d'établissement ne renvoient pas systématiquement de valeur pour ce champ. Le paramétrage des IDP reste hétérogène dans le cadre de la Fédération d'Identité et il n'existe actuellement que peu de règles communes de ce qui doit être renvoyé. Le support de shibboleth auprès des utilisateurs est également rendu difficile par l'absence de norme dans la fédération, pour identifier une personne. De notre côté, la résolution des problèmes consiste bien souvent à fournir un compte et rediriger l'utilisateur vers notre IDP.

2.5. Phase de déploiement

2.5.1. Organisation du déploiement

Le déploiement dans les laboratoires est piloté par les Délégués Régionaux de l'Inserm et les équipes informatiques régionales. Cette stratégie régionale a permis de faciliter le déploiement et la communication tant en interne qu'auprès de nos partenaires.

Le cahier électronique n'est plus individuel et s'organise à l'Inserm autour de groupes fonctionnels dépendants des périmètres de recherche, plus que des périmètres administratifs. Le cahier CLÉ va correspondre à un groupe de recherche travaillant ensemble sur des projets communs, autour d'un responsable (PI). Nous créons également des cahiers pour les plateaux techniques et les contrats industriels. Chaque cahier est indépendant. Il existe cependant des possibilités de partager certains éléments tels que des protocoles, des équipements et des inventaires.

Dans chaque UMR, le déploiement est réalisé sous l'autorité du Directeur. C'est lui qui va demander la création des cahiers, après avoir défini l'organisation interne, le nombre total de cahiers nécessaires, et désigner pour chacun d'entre eux le responsable (PI).

Voici les étapes du déploiement dans une UMR:

- inscription – validation ;
- réunion préparatoire (présentation et d'étude d'implantation) ;
- demande – ouverture de cahiers ;
- réunion de lancement dans le laboratoire ;
- formation – support des utilisateurs.

2.5.2. Documentation

Outre la documentation technique, la rédaction d'une documentation complète à destination des différents acteurs et notamment des utilisateurs finaux a été réalisée très largement en amont du déploiement. Une partie de ces documents a été rédigée en collaboration avec le Réseau Inserm Qualité (RIQ). L'ensemble des documents est disponible en langue française et anglaise. Parmi les principaux documents réalisés, nous pouvons citer:

- des présentations PowerPoint ;
- des vidéos de présentation ;
- une FAQ ;
- de la documentation :
 - « Conseils d'utilisation » (RIQ) ;
 - « Check-list avant d'utiliser CLÉ » (RIQ) ;
 - modes d'emploi pour les administrateurs ;
 - argumentaire pour les partenaires institutionnels et industriels ;
 - formulaires.

2.5.3. Formation et support des utilisateurs

CLÉ inclut une plateforme d'e-learning. Elle est accessible en permanence et de manière illimitée à tous les utilisateurs connectés et permet grâce à de courtes vidéos de démonstration et des quizz, d'appréhender rapidement les différentes étapes d'utilisation.

Des webinars permettent de compléter la formation sur des sujets particuliers comme l'administration du cahier ou la constitution des inventaires.

Un tchat en ligne, intégré à l'application, permet aux utilisateurs de dialoguer avec l'éditeur et d'obtenir rapidement des réponses à leurs problématiques fonctionnelles.

2.5.4. Communication

La communication a été réalisée par le Département de l'Information Scientifique et Communication via différents canaux. Une première présentation du CLÉ a été faite dans la lettre d'information de l'Institut sous la forme d'un article avec une interview des chefs de projet. Elle a été complétée par la mise en ligne d'une vidéo de présentation réalisée dans 2 laboratoires pilotes (format court de 3 min). Cette campagne a été relayée ensuite sur l'Intranet où une rubrique dédiée regroupe toutes des informations et les documents à l'usage des utilisateurs.

2.5.5. Bilan provisoire

Depuis décembre 2018, nous avons un déploiement important mais progressif de la solution dans les laboratoires. Même si l'attrait de CLÉ est fort au départ, nous constatons ensuite une certaine inertie entre la demande initiale et sa mise en œuvre. C'est notamment le cas au moment des choix d'organisations (nombre de cahiers, désignation des responsables) puis au niveau de chaque groupe de recherche au moment d'accéder, de paramétrer, de s'organiser et de définir les rôles.

Pour les 350 structures de recherche que compte l'Inserm, presque la moitié ont lancé la démarche. Nous avons 10 mois après le lancement, un peu plus de 400 cahiers ouverts et 2300 utilisateurs.

3. Nouveaux enjeux pour les ELN

3.1. Opportunités de jouer un rôle fédérateur

L'ELN peut avoir un rôle fédérateur dans le SI du laboratoire et devenir un LIMS (Laboratory Information Management System) complet. Il peut répondre aux besoins des utilisateurs de s'interfacer avec des bases de données et d'autres applications du laboratoire. Il peut également aller plus loin et s'interfacer plus largement avec un catalogue institutionnel de services, tant scientifiques qu'administratifs.

Les exemples d'interfaces indiqués dans le schéma ci-dessous ne sont pas forcément déjà mis en œuvre mais proviennent de demandes déjà évoquées à l'Inserm. Nous avons :

- dans le domaine collaboratif, la gestion des agendas, des emails et des SMS ;
- dans le domaine administratif, une interface avec l'outil de gestion financière dans le cadre des achats ,une interface avec l'outil de gestion de la feuille de temps des projets de recherche, ainsi qu'une interface avec l'outil de gestion des risques dans le cadre de la gestion des inventaires du laboratoire ;
- dans le domaine scientifique, l'interface avec l'application d'accès aux ressources bibliographiques ;

- dans le domaine de la donnée, la prise en compte des aspects de stockage des données brutes, de GED (Gestion Électronique de Documents), de traitement des données, d'entrepôts de données et d'archivage.

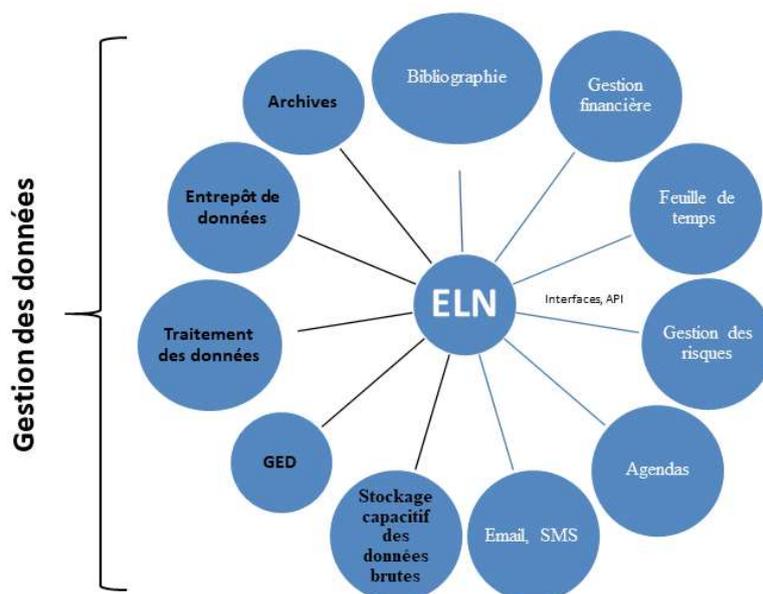


Figure 5 - L'ELN au cœur du SI du laboratoire

3.2. Interfaces de CLÉ : travailler avec des API

La solution Labguru choisie à l'Inserm offre la possibilité de travailler avec des API (Application Programming Interface). Les utilisateurs eux-mêmes peuvent développer des interfaces avec leurs propres applications, pour échanger des données avec leur cahier de laboratoire. Ces API basées sur JSON / REST, sont publiées sur <https://my.labguru.com/api/docs>. Pour la communauté, des ressources additionnelles sont disponibles sur Github <https://github.com/BioData/labguru-api-examples> (par exemple : Python wrapper of Labguru API) et une aide peut être apportée sur Slack (Labgurus channel labgurus.slack.com).

De nouvelles applications vont donc pouvoir interagir avec CLÉ, envoyer des requêtes, recevoir des données. Via les API, il est possible de consulter, créer, modifier les principaux éléments d'un cahier, à savoir un projet de recherche, une expérience, un protocole, un élément en stock, un équipement, un document annexe, ou même des éléments plus spécifiques comme le DataSet (jeu de données) ou le SOPs (Standard Operating Procedure) utilisés notamment dans la recherche clinique.

Nous pouvons donner comme exemple le plus abouti, la possibilité offerte aux équipes de recherche de travailler sur leur cahier depuis l'outil de statistique R (<https://www.r-project.org/>). Le script LabguruR, mis à disposition des utilisateurs (<https://github.com/BioData/LabguruR>) permet de travailler directement depuis la console R sur les données expérimentales (Dataset) stockées dans Labguru et d'enregistrer directement les résultats et graphiques obtenus.

3.3. Solution pour répondre aux enjeux du moment

Il nous semble que les ELN sont en capacité d'apporter des solutions pour répondre à deux enjeux majeurs actuels pour la recherche, à savoir l'amélioration de la gestion de la donnée et la reproductibilité des expériences. Ces sujets sont abordés notamment dans la Charte nationale de déontologie des métiers de la recherche [2], signée en 2015 par les acteurs français de la recherche publique et la Déclaration de Singapour sur l'intégrité en recherche [3]. Même si l'ELN ne fait clairement pas partie de l'OpenData, il se situe en amont et va permettre de fournir dans le cadre du DMP (Data Management Plan) une solution pour retrouver et décrire la manière dont seront obtenues, documentées, analysées les données produites au cours d'un projet de recherche. Le projet européen H2020 [4] impose désormais dans certains cas l'usage du DMP, pour garantir la production de données de recherche FAIR (Findable, Accessible, Interoperable, Reusable).

Pour améliorer la gestion de la donnée, les ELN devraient se doter, de toutes les fonctionnalités d'une solution de GED (Gestion Électronique de Documents). Certains laboratoires ont fait le choix inverse, celui d'utiliser une solution de GED comme cahier de laboratoire. Une solution de GED générique sera cependant moins spécifique. Elle aura sans doute besoin de beaucoup de paramétrages et rencontrera des difficultés pour associer des bases de données, par exemple d'inventaires.

L'amélioration de la reproductibilité des expériences passe par la conservation du descriptif et de l'ensemble des éléments associés. Ceci permet de garantir que l'expérience a été faite dans les mêmes conditions et que les démarches de mesures sont strictes. La conservation des données brutes, de l'ensemble des résultats, des moyens mis en œuvre (par exemple le protocole, le logiciel et sa version utilisée...) est un enjeu majeur pour améliorer la traçabilité et la reproductibilité des expériences. Il est essentiel que les ELN intègrent rapidement le besoin de stockage capacitif. Ce besoin s'exprime en pétaoctets et se heurte techniquement pour l'instant à des problématiques réseau : comment, en effet, peut-on récupérer et centraliser les données brutes, le plus souvent générées sur les sites lors des phases d'acquisition et de traitement ?

La réponse actuelle de la solution Labguru passe par la mise en place d'un agent de réplication des données sur les postes d'acquisition ou de traitement, mais aussi par des possibilités d'interfaces directes via l'utilisation d'API. Il sera également important pour atteindre l'objectif de pouvoir conserver les programmes dans la version ayant participé au traitement des données.

Le Big data, le traitement voire l'exploration des données s'implantent très rapidement dans le laboratoire, bousculent les habitudes et imposent une vraie rupture dans la gestion de son informatique. Les ELN commencent à s'adapter et n'ont pas fini d'évoluer. Les API ouvrent en tout cas de nouvelles perspectives aux utilisateurs.

Bibliographie

- [1] Dupré PG, Brizzi F. Expérimentation du cahier de laboratoire électronique à l’Inserm : les apports de l’électronique au cahier de laboratoire. Dans Actes du congrès JRES2017, Nantes, Novembre 2017 ; https://conf-ng.jres.org/2017/document_revision_2866.html?download.
- [2] Charte française de déontologie des métiers de la recherche, Janvier 2015 ;https://www.hceres.fr/sites/default/files/media/downloads/2015_Charte_nationale_deontologie_190613.pdf.
- [3] Déclaration de Singapour sur l’Intégrité en recherche, Juillet 2010 ; https://www.jsps.go.jp/english/e-kousei/data/singapore_statement_EN.pdf.
- [4] Projets H2020: Données de recherche FAIR ; <http://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html>.