LE PATRIMOINE DES DONNEES DE LA RADIOACTIVITE DANS L'ENVIRONNEMENT : COLLECTE, CAPITALISATION, SCRUTATION ET VALORISATION

Olivier PIERRARD, Miriam BASSO, Marjorie LELIEVRE, Jean-Pierre BENOIT, Maxime MORIN

IRSN
B.P. 17 - 92262 Fontenay-aux-Roses Cedex
olivier.pierrard@irsn.fr

Les analyses radiologiques réalisées chaque année sur des échantillons environnementaux par les différents laboratoires français produisent des dizaines de milliers de résultats. Plusieurs milliers de ces données concernent en particuliers les denrées et bioindicateurs qui sont régulièrement publiées sur le site internet du Réseau National de Mesure de la radioactivité dans l'environnement (ou RNM, www.mesure-radioactivite.fr) dont l'IRSN assure la gestion.

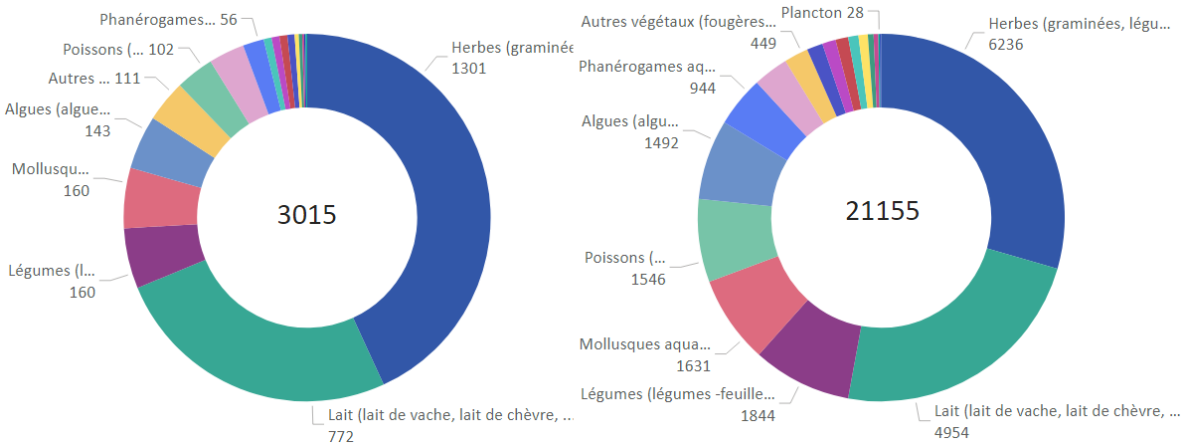

Figure 1 : nombre de prélèvements (à gauche) et de mesures (à droite) relatifs aux denrées et bioindicateurs publiés au RNM en 2020 (40000 environ depuis 2010)

Pour parvenir jusqu'au site internet du RNM, les informations tirées de la mesure de la radioactivité dans l'environnement doivent parcourir un cheminement complexe semé de processus numériques visant l'acquisition, le stockage, les conversions et les contrôles qui permettront de garantir leur qualité et donc leur exploitation correcte. Les producteurs de données doivent ainsi mettre en œuvre des moyens numériques en charge de garantir l'intégrité de ces données, du terrain jusqu'à la consultation.

L'acquisition de données de terrain constitue une étape fondamentale durant laquelle l'opérateur doit relever l'ensemble des informations contextuelles associées à la mesure ou au prélèvement et qui représentent autant de « clés » qui permettent de rechercher ou de représenter ultérieurement les résultats d'analyse ou de mesure dans les outils de restitution. Pour acquérir et enrichir les données liées aux analyses radiologiques d'échantillons, la plupart des laboratoires des producteurs dispose d'une application infomatique de gestion de l'information en laboratoire (ou LIMS) qui doit combiner plusieurs contraintes fonctionnelles : robustesse, souplesse d'utilisation (gestion de prélèvements et analyses programmés ou non), simplicité et ergonomie, compatibilité avec les systèmes amont (terrain) et aval (capitalisation, publication des données au RNM par exemple).

Véritable fil d'Ariane de l'information numérique entre la réception d'un échantillon et la validation des résultats d'analyses réalisées sur cet échantillon, le LIMS nécessite souvent un coûteux maintien en conditions opérationnelles imposé par la garantie d'un taux de disponibilité important tout en faisant face à l'obsolescence rapide de ses composants.

Les pratiques et les systèmes (dont les LIMS) évoluant régulièrement, la capitalisation des informations numériques dans le temps est un enjeu important pour la plupart des entités en charge d'études et de surveillance de la radioactivité dans l'environnement. Ainsi, les données numériques acquises au fil du temps représentent un patrimoine important qu'il convient de sauvegarder pour mieux exploiter les chroniques. Comme d'autres acteurs de la surveillance radiologique, l'IRSN a mis en place un système d'entreposage numérique (SYRACUSE) pour organiser le stockage des informations acquises par une multitude d'outils (base de données, fichiers ou LIMS) depuis plus d'une trentaine d'années. Pour organiser correctement les flux de données vers SYRACUSE, une branche dite de « modélisation » est développée spécifiquement pour chaque source. Elle intègre les règles de transformation des données (traduction selon des référentiels communs, enrichissement d'information manquantes, recalcul pour uniformisation) et de contrôle, garantes de la qualité et de l'unicité des données intégrées dans cet entrepôt.

Pour améliorer la qualité de son patrimoine de données relatives à la radioactivité dans l'environnement, l'IRSN développe également des outils d'analyse statistique de scrutation visant à identifier l'éventuelle présence d'anomalies ou valeurs aberrantes dans son entrepôt de données. Ces dernières peuvent influencer fortement la valeur des indicateurs statistiques utilisés régulièrement comme, par exemple, la moyenne calculée à partir d'un jeu de données « matrice-radionucleide-unité » pour définir un niveau de référence sur une période. Par conséquent, leur présence peut fausser la compréhension du jeu de données et amener à émettre des conclusions erronées. Il est essentiel de les identifier puis de les corriger ou de les expliquer (marquage connu, problème lié à la saisie dans la base de données, marquage à investiguer). Dans une première approche, l'IRSN a mis en œuvre un protocole de détection d'anomalies basé sur la méthode dite de « Tukey », une des méthodes statistiques de scrutation à une seule dimension parmi les plus classiques. Selon cette méthode, les données sont considérées comme « anormales » lorsque leur valeur dépasse un seuil calculé à partir des quartiles de la distribution des mesures considérées et/ou par rapport à des jeux de données de référence considérées représentatives du bruit de fond environnemental. Sont prévus les développements de nouveaux algorithmes permettant d'identifier, par exemple, les anomalies d'évolution temporelle (décrochages notamment) ou d'intégrer la dimension spatiale à l'analyse afin d'identifier des valeurs anormales par rapport à leur contexte géographique.

Enfin, l'IRSN expérimente, développe et maintient des interfaces permettant d'exposer ces données de la manière la plus efficace, mêlant filtres multiparamétriques et visuels interconnectés developpés via des outils de data visualisation. Pour faire évoluer les restitutions classiques de type rapport d'étude ou de surveillance, ces mêmes outils sont également testés pour la valorisation des données selon le mode du data storytelling.