ÉCOLE NATIONALE DES CHARTES

Pauline Breton-Chauvet

Méthodologie de mise en conformité avec les principes FAIR

Métadonnées et données de New AGLAE

Présentation

Le stage réalisé à l'Accélérateur Grand Louvre d'Analyse Élémentaire (New AGLAE) du 29 mars au 31 juillet 2021 a consisté en l'analyse de l'état des données et des métadonnées actuellement produites par l'accélérateur, ainsi qu'en la proposition d'une méthodologie d'application des principes FAIR (*Findable, Accessible, Interoperable, Reusable*). Cette réflexion s'est inscrite dans celle menée par le Centre de Restauration et de Recherche des Musées de France (C2RMF) sur la gestion à long terme des données numériques et la mutualisation de gisements jusqu'ici relativement cloisonnés. La conservation pérenne des données produites à AGLAE est ainsi un axe cardinal de ce travail, avec la mise en œuvre de la méthodologie OAIS dès la genèse du cycle de vie des (méta)données. Par ailleurs, le stage a coïncidé avec la fin de la phase d'investigation d'Euphrosyne, devenue start-up d'État du Ministère de la Culture en avril dernier. Totalement pensée pour les utilisateurs, Euphrosyne a vocation à offrir dans un premier temps aux chercheurs et aux utilisateurs de l'accélérateur – la « communauté AGLAE - un accès distant aux données expérimentales ainsi qu'à leur traitement. À moyen et long terme, elle doit ensuite permettre l'ouverture, l'interrogation et le partage des données dans le respect des principes FAIR et des droits de la propriété intellectuelle [1]. Dans une dynamique de réciprocité, le stage a contribué à alimenter la réflexion autour des modalités conceptuelles, organisationnelles et technologiques de sa première phase de déploiement, tout en bénéficiant des inspirations plurisdisciplinaires des différents acteurs d'Euphrosyne.

Les livrables techniques sont les suivants :

1. Un inventaire de l'état FAIR des (méta)données d'AGLAE, avec les cartographies de l'existant.
2. Une méthodologie en vue de leur FAIRisation et de leur pérennisation dans le cadre plus large de la politique de gestion et de conservation du C2RMF.
3. Un modèle conceptuel et sa traduction en modèle relationnel pour un jeu de données expérimentales en vue d'une implémentation dans Euphrosyne.
4. Un modèle d'archivage des données d' AGLAE en conformité avec la méthodologie OAIS.

1. Ces deux phases de déploiement correspondent respectivement à Euphrosyne-manip et Euphrosyne-data.

5. Un modèle en graphe construit à partir des référentiels retenus et de leur organisation en vue d'une introduction dans le web sémantique.

Chapitre 1

Inventaire de l'existant

Les premières semaines du stage ont été consacrées à la découverte et à l'apprivoisement des données d' AGLAE, avant de procéder à une cartographie de l'existant. Une immersion dans les expériences menées à AGLAE, nourrie par des échanges riches avec les utilisateurs et l'équipe de l'accélérateur, a permis de réaliser cet inventaire au plus près des usages actuels. Nous avons cartographié les différentes étapes et objets des flux de données par techniques ainsi que par types d'analyse, en nous inspirant partiellement de la modélisation UML (*Unified Modeling Language*) [1].

1.1 *Findable* : état des (méta)données

L'un des axes cardinaux des principes FAIR repose sur la capacité des données à être facilement (re)trouvées tant par l'humain que par la machine. Cette exigence, traduite par le terme anglais *Findable*, implique le recours à un identifiant unique et stable, ainsi qu'une description avec des métadonnées riches et appropriées, enregistrées ou indexées dans un dispositif permettant de les interroger par l'intermédiaire d'un moteur de recherche.

Les ressources numériques d'AGLAE ne disposent pas à ce jour d'identifiant autre que le nommage de fichier, instable et parfois répété, donc non conforme avec le principe d'unicité. Concernant les métadonnées, cette étape de travail a révélé un processus de transition informationnelle déjà en cours, avec de nettes inégalités de renseignement entre les différents types de fichier et, dans une moindre mesure, entre les différentes techniques d'analyse, celles particulièrement utilisées comme PIXE et PIGE ayant bénéficié d'un soin particulier. Par ailleurs, nous avons constaté des redondances dues à la variété des supports de métadonnées. Ainsi, les informations renseignées dans le cahier de laboratoire imprimé reprennent certains éléments figurant sur la demande de temps de faisceau tels que le nom et la responsabilité de l'analyste, la technique d'analyse souhaitée, l'objet ou le lot ciblé par l'expérience. Y figurent également les détecteurs choisis, les filtres et les

1. Se reporter au document « Cartographie_AGLAE » contenu dans le répertoire général des livrables techniques.

standards sélectionnés, autant d'informations réitérées dans le fichier Excel généré automatiquement et valable pour toute l'expérience. Toutefois, ce cahier détient également des informations uniques et non dématérialisées sur certaines anomalies ou corrections expérimentales, sans lesquelles il est impossible de comprendre la manipulation et le traitement des données qui ont conduit à leur état définitif. Les informations du contexte expérimental existent donc mais ne sont pas organisées de façon à offrir une intelligibilité pérenne au niveau de granularité approprié. Par ailleurs, certains utilisateurs renseignent uniquement les champs référencés du cahier de laboratoire, sans utiliser l'espace libre de rédaction, et réservent les commentaires inhérents à la chaîne expérimentale à des supports personnels qu'ils conservent et sur lesquels l'équipe d'AGLAE n'a donc pas la main. Il n'est ainsi pas rare de devoir faire appel à la mémoire des analystes et des chercheurs présents lors d'une expérience passée lorsqu'il y a nécessité d'en comprendre les résultats. De plus, il n'existe pas encore de jeu de données tel que nous l'avons récemment défini mais un regroupement de fichiers par nom d'utilisateur et date de manipulation. L'accès à des données expérimentales précises s'en trouve par conséquent fortement entravé si le cahier de laboratoire n'offre pas d'informations contextuelles complètes et satisfaisantes.. La compréhension et la possibilité du rejeu de certaines expériences sont donc, à l'heure actuelle, l'apanage exclusif de l'expérimentateur et de l'éventuel groupe de chercheurs qui l'accompagne. Nous constatons ici un revers possible de l'autonomie de certains utilisateurs, tentés de s'émanciper d'AGLAE pour la gestion et le stockage des données et des métadonnées expérimentales, tant pour des raisons pratiques d'accès qu'en vertu d'une appropriation intellectuelle. En plus de ces informations textuelles et chiffrées, de nombreux utilisateurs prennent également des photographies des zones d'intérêt analysées au cours de l'expérience, qui constituent des métadonnées singulières mais précieuses, échappant à la conservation et au stockage d'AGLAE.

Parallèlement aux métadonnées renseignées manuellement par les utilisateurs, il y a celles récupérées et versées automatiquement dans l'en-tête et dans le nommage de la plupart des fichiers à l'aide d'un programme développé sur mesure à AGLAE, en vue de la compréhension rétrospective des données. Il s'agit alors moins d'une volonté de pérennisation dans une perspective d'ouverture que d'améliorer l'intelligibilité des données par les utilisateurs et l'équipe d'AGLAE. Ces deux paradigmes se rejoignent toutefois dans les évolutions de pratique et de structure informationnelles induites. Que l'on soit dans le cadre de la méthode d'analyse PIXE ou PIGE, nous retrouvons des champs de métadonnées similaires par type de fichier. Les fichiers spectres de l'analyse ponctuelle en format ASCII contiennent tous une en-tête de deux lignes : la première indique le nombre de canaux du spectre, la seconde répète certaines informations du contexte expérimental figurant à la fois dans le fichier Excel général et dans le cahier de laboratoire. S'y ajoutent des métadonnées à la granularité plus fine - bien que partiellement appropriée - telles que la somme du spectre ou le temps d'acquisition. Les en-têtes des fichiers LST de l'imagerie

se décomposent également en deux lignes : la première relative à la dimension ainsi qu'à la résolution de la cartographie ; la seconde réitérant les informations générales du contexte expérimental. Cette redondance est également présente dans les fichiers EDF résultant de la conversion des fichiers LST. Elle témoigne d'une capillarité contextuelle nécessaire lorsqu'un jeu de données est organisé en fonction d'une arborescence d'usage et non de hiérarchisation sémantique de contenus.

Certaines limites technologiques placent certaines données hors de leur environnement informationnel et contextuel. En effet, les métadonnées d'un fichier LST structurées dans une en-tête ne peuvent être lues par le convertisseur et ne sont donc pas incorporées automatiquement au fichier EDF de sortie. DATAFURNACE, le logiciel de traitement de certaines données RBS ne parvient pas non plus à traiter de fichiers comportant une en-tête. Pour ces cas spécifiques, les seules métadonnées sont celles réinjectées a posteriori dans les fichiers issus de la conversion, et/ou celles intégrées dans le nommage des fichiers, ce dernier n'étant pas normalisé dans le cas de la méthode RBS. Qu'il soit chargé de pallier à certains écueils technologiques ou à améliorer l'aisance d'accès et de recherche pour les utilisateurs, le nommage a acquis une dimension prédominante dans la signalisation et l'organisation des données d'AGLAE. Les logiciels de conversion et de traitement dépendent d'ailleurs de lui pour la sélection de paramètres et d'éléments spécifiques. À titre d'exemple, les fichiers .par générés par GUPIXWin puis traités par TRAUPIXE doivent forcément se terminer, dans leur nommage, par « _nom du détecteur ». Le lien de chaque fichier .par au détecteur adéquat est indispensable pour TRAUPIXE qui le traite nécessairement en fonction du/des détecteur(s) sélectionné(s) [2]. Ce système d'intégration de métadonnées au nommage, déployé faute d'un espace approprié d'indexation et de renseignement, entraîne l'élaboration de noms de fichiers parfois excessivement longs et parfois délétères pour leur accessibilité. Le nom des fichiers .png – ou .dat si le contenu est en format ASCII - pour la fin de flux de l'imagerie PIXE ou PIGE est ainsi composé : de la date, du numéro d'analyse, de la référence du projet, du nom du détecteur et d'un éventuel commentaire de l'utilisateur. Quant aux fichiers spectres primaires, leur nom résulte de la combinaison de la date, du numéro d'analyse, de la référence de l'objet, du nom du projet suivi de « _IBA.x ». Cette structuration composite peut être problématique selon les points d'accès recherchés, variables selon les utilisateurs, lesquels souhaitent parfois effectuer une recherche sur le dernier élément renseigné dans le nom. Les noms des fichiers TRAUPIXE en offrent un exemple éloquent. Ils se composent de la date, de la matrice avec le détecteur associé à la matrice ou à la trace et se terminent par l'indication de l'élément pivot. Ce dernier correspond à un élément chimique choisi par l'utilisateur, sous réserve d'être présent dans les différents détecteurs PIXE et en quantité suffisante. Extrêmement important pour le traitement et l'interprétation des données PIXE, il peut

2. En plus des détecteurs réels d'AGLAE, il existe des détecteurs qualifiés de virtuels qui correspondent à la combinaison de plusieurs détecteurs réels.

faire l'objet d'une recherche ciblée dans des données historiques. Son apparition en fin de nommage constitue en cela un obstacle car seul le début du nom de fichier est directement visualisable.

Il n'existe pas de règle commune ou même de bonne pratique visant à harmoniser les dénominations, laissées à la libre appréciation des utilisateurs. Nous sommes désormais dans une transition hybride où la gestion des fichiers répond à la fois aux besoins pratiques des analystes et à un effort documentaire récent en vue de l'accès et de l'ouverture des données historiques. Par leur organisation, les données sont ainsi toujours reliées de façon prépondérante aux utilisateurs. En revanche, la dénomination actuelle des fichiers traduit une volonté de restitution du flux de données dans ses composantes technologiques et scientifiques, sans présomption des usages qui en seront faits. À titre d'exemple, les fichiers générés par TRAUPIXE, qui ne faisaient pas initialement l'objet d'un nommage standardisé, ont ensuite été préfixés par GUPIX, suscitant une ambiguïté avec les fichiers issus de GUPIXWin. TRAUPIXE s'est ainsi finalement substitué à ce choix, mais sans entraîner de modifications rétrospectives sur les noms de fichiers antérieurs à cette syntaxe. Les fichiers 2D-EDF ont également connu plusieurs versions de nommage, et sont désormais regroupés par répertoire utilisateur comprenant des répertoires par année, eux-mêmes conteneurs de répertoires par run. Leurs noms se terminent maintenant systématiquement par « IBA- nom du détecteur_EDF ». Il faut souligner le fait que chacune de ces modifications a ensuite dû être appropriée par des analystes souvent soucieux de conserver des repères familiers lors de l'accès et du traitement des données. Par ailleurs, bien qu'incités à respecter les bonnes pratiques de nommage mises en place à AGLAE, les utilisateurs demeurent libres d'influer sur la sémiologie des données qu'ils contribuent à produire.

Notre inventaire de l'existant a mis en relief une abondance de métadonnées, dont certaines sont susceptibles d'être réorganisées sans affecter les utilisateurs, puisque inusitées. Insistons sur le fait qu'il s'agit bien plutôt d'un réajustement que d'une somme d'ajouts et de suppressions, car les métadonnées renseignées actuellement ont toutes leur utilité lors d'une ou plusieurs étape(s) du cycle de vie des données. Elles peuvent participer à la lecture diachronique de l'expérience et/ou à sa viabilité elle-même, tout en offrant chacune une pierre à l'édifice de la pérennisation. Dans leur organisation actuelle, les métadonnées associées aux différents fichiers contenus dans un répertoire utilisateur ne fournissent cependant qu'une intelligibilité partielle des données dans les étapes successives de leur cycle de vie. De réels efforts pédagogiques sont déployés auprès des chercheurs par l'équipe d' AGLAE pour les accompagner dans l'assimilation de ce qui représente un véritable changement de culture. En effet, les utilisateurs de l'accélérateur n'ont jamais, pour la plupart d'entre eux, envisagé que la matière première de leurs publications puisse être rendue accessible à un public élargi, ou même entièrement ouverte.

Quant au dispositif d'enregistrement et/ou d'indexation des (méta) données, il re-

pose sur une structure informatique exclusivement présente sur le site de l'accélérateur. Un programme présent sur l'ordinateur d'acquisition enregistre les données dans des fichiers LST copiés automatiquement sur le disque dur à la fin de la manipulation. Le reste des données, notamment les fichiers spectres, les photographies des zones d'intérêt ainsi que les rapports Excel sont enregistrés dans un serveur de stockage en réseau (NAS), dans un répertoire sélectionné en début de manipulation contenu par « C2RMF Users ». Cette distinction est due à la lenteur d'enregistrement et aux limites techniques du NAS initial, suscitant une concurrence entre les besoins de rapidité d'accès et de lecture et les structures nécessaires à l'archivage.

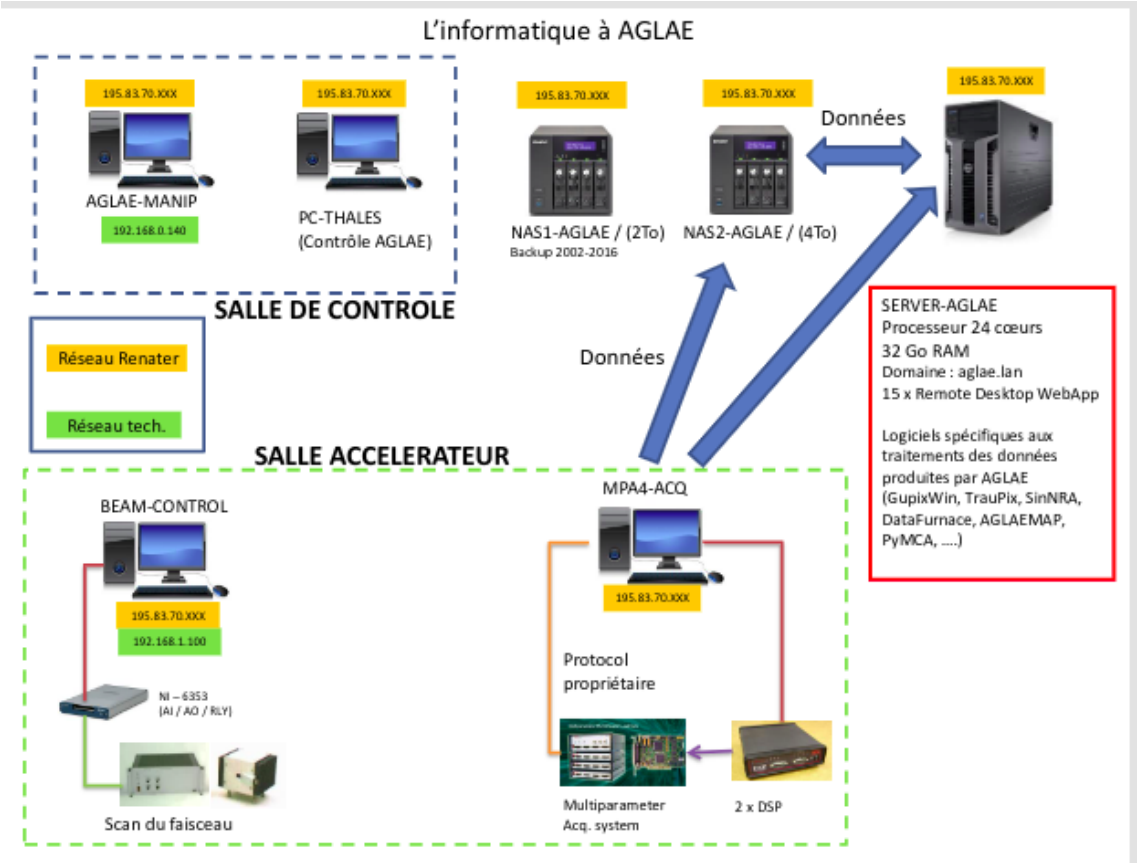


FIGURE 1.1 – Présentation de l'infrastructure de stockage des données de New AGLAE @Laurent Pichon

Le NAS2 dédié à l'enregistrement et au stockage des données d'AGLAE bénéficie d'un fonctionnement en RAID [3] avec des duplications sur plusieurs disques durs. Cela permet d'éviter la perte de données en cas de panne définitive de l'un de ces derniers. Néanmoins, leur regroupement au sein de la même machine, située elle-même à l'endroit de production des contenus, les expose simultanément aux risques liés à un accident ou à un événement imprévisible susceptible de causer des dégâts. Aucune duplication externalisée n'existe, si ce n'est sous forme empirique, partielle et fragmentée par l'intermédiaire

3. RAID, acronyme de *Redundant Array of Independent Disks*, soit « matrice redondante de disques indépendants » désigne une technologie permettant de stocker des données sur de multiples disques durs.

des multiples disques durs personnels des utilisateurs qui participent aux expériences d'AGLAE.

Par ailleurs, il n'existe pas encore de silo commun à l'ensemble du laboratoire de recherche du C2RMF. Les (méta) données produites par AGLAE ne sont donc pas mutualisées à celles des autres techniques scientifiques pratiquées sur un objet ou un lot concerné par une demande de service ou un projet de recherche transversal. Cette absence de mutualisation induit l'absence de dispositif dédié spécifiquement au stockage et au requêtage des informations documentaires et conceptuelles sur les contenus conservés.

1.2 *Accessible* : état de l'accessibilité des données d'AGLAE

Le principe d'accès promu par le credo FAIR n'induit pas nécessairement l'ouverture totale des données. Néanmoins, celles-ci doivent être récupérables par leur identifiant en utilisant un protocole standard de communication ouvert, libre et d'usage universel. Les (méta) données doivent être disponibles à des conditions connues grâce à des licences claires et visibles telles que *Creative Commons* pour ne citer qu'elle. Si les données ne sont pas ou plus disponibles, les métadonnées doivent demeurer cependant accessibles. En cas de fermeture – même partielle – des données, il faut en indiquer la durée et la raison. De plus, si des outils logiciels de visualisation et/ou de traitement particuliers sont nécessaires pour accéder aux données et les exploiter, ils doivent être documentés de façon ouverte et explicite. L'utilisation de logiciels libres est en cela particulièrement recommandée. Enfin, le lieu de consultation des données doit être clairement identifié et accessible.

Les données d' AGLAE, identifiées uniquement par le nommage des fichiers qui les contiennent, ne sont pas encore requêtables et récupérables par un protocole standard de communication, quel qu'il soit. Elles nécessitent une navigation « manuelle » parmi l'arborescence de répertoires et de fichiers, par un utilisateur ayant connaissance au préalable de l'architecture et de l'organisation des données. De surcroît, les utilisateurs d' AGLAE, pour la plupart chercheurs et ingénieurs en sciences des matériaux patrimoniaux, se livrent toujours à des pratiques d'embargo durant le délai antérieur à la publication [4], malgré l'absence de socle juridique pour légitimer la fermeture des contenus « formellement achevés » [5] ne contrevenant pas au RGPD [6] ou à certaines réserves telles que le secret bancaire ou fiscal. Toute démarche de FAIRisation doit prendre en compte les motivations de cette rétention, sous peine d'aller à l'encontre des utilisateurs d' AGLAE. Du point de vue de

4. Ce délai est aujourd'hui approximativement de trois ans.

5. Entretien du 3 juin 2021 avec Laurent Romary, directeur de recherche à l'INRIA et président de l'ISO/TC 37, comité technique de l'Organisation internationale de normalisation.

6. *Règlement général sur la protection des données*, publié le 27 avril 2016 par le Conseil européen.

la plupart d'entre eux, l'appropriation à travers l'attribution silencieuse d'une propriété intellectuelle au sens moral commence dès la définition et la sélection de paramètres nécessaires à la découverte et à la révélation inhérentes à l'objet étudié. Cela explique la fréquence des pratiques de rétention des données par les utilisateurs d'AGLAE, craignant de s'en voir dérober la paternité. À l'instar de l'archéologue revendiquant la responsabilité de la découverte d'un objet enfoui depuis des millénaires, le chercheur est attaché de façon presque affective aux informations qu'il a exhumées au moyen d'un complexe outillage physique, technique et numérique, d'une expertise et d'une formation intellectuelle exigeantes, et d'un protocole d'acheminement des œuvres souvent lourd et contraignant. La singularité de cette chaîne, qu'il est souvent impossible de répéter tant elle est coûteuse et complexe, suffit à justifier le sentiment de la *trouvaille* et la perception de sa préciosité, une fois l'expérience terminée.

Les analystes et chercheurs d' AGLAE redoutent également le pillage à des fins de réalisation de contrefaçons. En effet, les résultats des analyses réalisées par l'accélérateur offrent indirectement, aux esprits mal intentionnés, la recette chimique pour produire un faux dont la composition sera parfaitement similaire à celle d' un objet authentique. Le législateur ne s'est pas encore penché sur ce risque, auquel les producteurs et utilisateurs des données d' AGLAE répondent actuellement par une défiance à l'égard de toute forme de diffusion et d'ouverture autre que l'interprétation définitive sous forme éditoriale. Ni les données ni les métadonnées ne sont aujourd'hui identifiées et accessibles pour des individus extérieurs à AGLAE. Il n'existe donc nulle part d'indication de lieu de consultation, qui correspond exclusivement à celui de production des données. Nous en sommes encore à une gestion exclusivement orientée vers les utilisateurs actuels, avec de fortes difficultés d'accès aux données historiques. En effet, si les serveur et NAS d' AGLAE stockent les données brutes et intermédiaires, la plupart des utilisateurs gardent la main sur les données traitées et interprétées, conservées sur un disque dur ou une clé personnels. De vastes pans de données sont ainsi totalement inaccessibles et ne peuvent bénéficier d'une réorganisation et d'une gestion rétroactives en vue d'une éventuelle FAIRisation.

L'accessibilité des données pose naturellement la question de la propriété des logiciels, AGLAE utilisant un éventail hétérogène où les logiciels "maison" conçus en interne côtoient et sont même parfois interdépendants de logiciels propriétaires payants. Chaque groupe de données par technique d'analyse (PIGE, PIXE, RBS ou IBIL) est pris en charge par un outil développé par l'équipe d'AGLAE pour ses usages particuliers et faisant appel à un moteur de calcul spécifique. Les données PIXE sont ainsi traitées par le logiciel « maison » TRAUPIXE qui fait appel au moteur de calcul GUPIXWin. Dans le cadre de l'extraction et du traitement quantitatif, les données converties par les logiciels *ad hoc* sont ensuite extraites en format EDF (*ESRF Data Format*) permettant leur manipulation par des outils en accès libre tels que PyMCA ou des logiciels « maison » de traitement et/ou de visualisation des données. Les logiciels payants et propriétaires utilisés tels que

GUPIX ou DATAFURNACE sont clairement documentés, ce qui n'est pas le cas de ceux développés à AGLAE.

1.3 *Interoperable* : état de l'interopérabilité

L'étude de l'état actuel des données d' AGLAE au regard de l'exigence d'interopérabilité est le troisième axe développé au cours de notre inventaire. L'interopérabilité renvoie à la capacité d'un système informatique à fonctionner avec d'autres systèmes, existants ou futurs, sans restriction de mise en œuvre ou d'accès. À la fois syntaxique et sémantique, elle implique une compatibilité tant conceptuelle qu'informatique. Elle impose de recourir à des contenus et des formats conformes aux grands standards internationaux (notamment ceux portés par le W3C), ainsi qu'à des métadonnées contextuelles précises.

Les (méta) données d' AGLAE ne sont pas encore organisées, structurées, identifiées et formatées de façon à les mettre en relation avec des ressources internes ou externes au C2RMF. Simplement stockées dans des répertoires d'utilisateurs et non dans une base de données structurée, elles ne sont pas sous-tendues par une architecture conceptuelle ou sémantique qui favoriserait leur extraction et leur partage avec d'autres systèmes. Leurs formats sont par ailleurs hétérogènes : ASCII, Excel, JPEG, CSV, format binaire, etc. La norme JPEG (*Joint Photographic Experts Group*) est conforme aux standards répertoriés et recommandés pour l'enregistrement et l'algorithme de décodage des images. Quant au CSV (*Comma-separated values*), format informatique d'échange de données ouvert, il est sur le déclin et exclusivement préconisé pour les échanges entre application et utilisateur. Le standard XML est à privilégier pour tous les échanges entre applications ou systèmes n'impliquant pas d'utilisateurs [7]. Ainsi, nous constatons que la plupart des formats des données d'AGLAE ne sont pas conformes aux standards préconisés pour l'interopérabilité, ou qu'imparfaitement avec un risque accru d'obsolescence.

Comme nous l'avons vu précédemment, les métadonnées inventoriées figurent exclusivement dans le contenu même des données ou dans leur nommage et ne bénéficient pas d'une indexation et d'un renseignement contextuel à la syntaxe et au format compatibles avec les standards de structuration de données recommandés. Il n'y a donc aujourd'hui ni interopérabilité interne ni interopérabilité externe.

1.4 *Reusable* : état de la capacité de réutilisation

Le dernier principe fondateur de la « déontologie » FAIR, traduit en anglais par le terme *reusable*, est bâti sur la capacité de réutilisation des données, ce qui la rend inter-

7. Direction Interministérielle du Numérique et du Système d'Information et de Communication de l'État, Référentiel Général d'Interopérabilité. Standardiser, s'aligner et se focaliser pour échanger efficacement, version 2.0, décembre 2015, mis à jour en décembre 2020.

dépendante de l'interopérabilité dont elle reprend les lignes. Ainsi, elle doit être facilitée par l'utilisation de standards communs, grâce à des bases rassemblant des données claires, contrôlées et rigoureusement décrites. Les informations de provenance sont là encore indispensables, et sont tenues d'être présentées avec une licence d'utilisation compréhensible et accessible.

Les (méta) données d'AGLAE n'étant pas accessibles, si ce n'est par les utilisateurs, l'équipe et les familiers d'AGLAE, elles ne sont actuellement pas non plus réutilisables. Quelques informations indirectes inhérentes au contexte institutionnel, scientifique et administratif des analyses pratiquées sont néanmoins accessibles sous format PDF à travers les rapports définitifs établis pour le C2RMF par un responsable de projet de recherche ou de demande de service. Versés dans la base EROS (*European Research Open System*) du C2RMF, ils se voient attribuer des identifiants numériques.

Chapitre 2

Méthodologie de FAIRisation

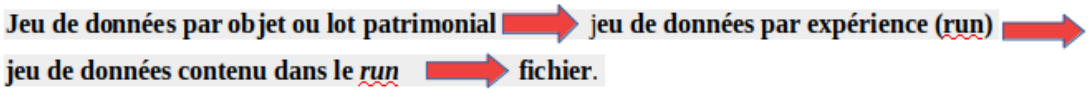
2.1 *Findable, accessible* : optimiser l'accès et la recherche des (méta)données

2.1.1 Identifier numériquement les ressources

La mise en conformité avec le principe *Findable* implique dans un premier temps pour les (méta) données d'AGLAE d'être identifiées de façon unique et d'être organisées selon une logique conforme aux différents usages envisagés. Afin de préparer leur versement dans un futur silo commun au C2RMF, nous recommandons d'abord le réemploi de la syntaxe actuellement utilisé par certaines techniques du laboratoire – notamment l'imagerie et la photographie - pour l'identification numérique des données, et qui se décompose ainsi : préfixe de la technique utilisée, date et séquence numérique unique. Précisons que ce choix, qui permettra de répondre directement aux besoins des utilisateurs courants et d'une agrégation interne, peut avoir pour revers l'enfermement des données dans un silo applicatif, et une difficulté de cohérence technique avec d'autres silos externes. C'est pourquoi nous préconisons, dans la mesure du possible, le recours à un second identifiant unique, mis en correspondance numérique avec le premier. Cet autre identifiant sera bâti sur le format ARK (*Archival Resource Key*), répandu dans le monde des institutions culturelles publiques. ARK est favorisé car il confère une relative liberté à l'autorité nommante, bénéfique à l'établissement d'une politique de gestion adaptée à la complexité et à l'hétérogénéité des données d'AGLAE. En revanche, l'usage de ce format entraînera des missions supplémentaires en interne de gestion et de maintenance des identifiants, dont les coûts financiers et humains restent à définir. Construit sur la norme URI (*Uniform Resource Identifier*), il permettra l'introduction des données d'AGLAE dans l'architecture du Web, constituant l'une des clés de voûte de l'interopérabilité externe. Il est conseillé de générer des noms ARK opaques afin de garantir leur constance dans le temps. L'outil de gestion d'identifiants développé devra :

1. Générer un nom ARK conforme aux règles de nommage choisies par l'organisation.
2. Attribuer un identifiant en associant un nom ARK à une ressource en enregistrant ses métadonnées et son URL d'accès.
3. Interpréter l'identifiant lors de son appel en redirigeant vers la localisation actuelle de la ressource.
4. Fournir des métadonnées d'identification et une déclaration de permanence pour chaque ressource.
5. Conserver la trace de l'identifiant et de ses métadonnées même lorsque la ressource sera supprimée ou indisponible.

Ces identifiants ARK seront attribués aux ressources correspondant à chaque niveau de l'arborescence jusqu' au fichier contenant les données : Il apparaît inutile et inopportun



d'aller plus loin en terme de granularité. Précisons que le jeu de données par objet ou lot patrimonial n'existe pas encore et dépend de la construction d'une gestion mutualisée des données du laboratoire du C2RMF - en terme de stockage, d'identification, d'agrégation et d'interrogation – actuellement en cours. La mise en place d'un serveur commun, fer de lance de la mutualisation, de son système d'accès et de stockage partagé sera incontournable et se trouve déjà à l'étude.

L'usage et le renvoi à un DOI (*Digital object identifier*) [1] semble quant à lui approprié et particulièrement pertinent pour l'identification et le lien avec les données publiées, en raison de son utilisation par les plateformes plurisdisciplinaires d'archives ouvertes telles que HAL [2], destinées au dépôt ainsi qu'à la diffusion d'articles scientifiques de niveau recherche, publiés ou non. Un identifiant DOI est composé de deux parties séparées par un slash : un préfixe identifiant l'organisme éditeur et un suffixe identifiant l'objet chez ce dernier. Sa gestion dépend d'une Fondation internationale, responsable de l'organisation générale et de l'administration du système DOI [3].

2.1.2 Réorganiser, restructurer et enrichir les métadonnées

L'exigence de repérabilité des ressources d'AGLAE impose également de recourir à des métadonnées riches et appropriées. Dans une étroite association avec le processus de

1. Pour attribuer des DOI, il est nécessaire de s'enregistrer auprès d'une agence dédiée (DataCite, CrossRef ou l'INIST-CNRS en France) qui attribue à l'organisation un préfixe unique permettant d'attribuer un certain nombre d'identifiants.

2. HAL est une plateforme en ligne développée en 2001 par le Centre pour la communication scientifique directe du CNRS.

3. Une présentation synthétique est disponible au lien suivant : `https://bbf.enssib.fr/consulter/bbf-1998-03-0049-007`

transformation archivistique, ces dernières seront réorganisées et enrichies de métadonnées techniques, descriptives, de structure et de provenance. Au préalable, il sera indispensable de refondre les données historiques et d'organiser celles à venir en fonction du périmètre récemment défini pour un jeu de données expérimentales, contenant lui-même des jeux regroupés par méthode d'analyse.

Dans le cadre de la construction de la première phase de déploiement d'Euphrosyne, exclusivement dédiée au premier cercle d'utilisateurs d'AGLAE, l'équipe pluridisciplinaire du projet a dû en effet circonscrire les limites d'un *jeu de données* de façon à l'encapsuler dans une unité sémiologique cohérente, lourde d'enjeux pour la structure numérique qui la sous-tend. Cette étape génétique d'Euphrosyne a été appréhendée de façon collective et itérative, chaque proposition étant soumise au débat et aux éventuelles reprises des membres de l'équipe. Le fil d'Ariane de cette réflexion a été tissé dans une subjectivité prospective : toute personne souhaitant « rejouer » l'expérience pratiquée sur un objet patrimonial dans un futur proche comme lointain doit pouvoir le faire à partir de l'ensemble de données et d'informations mis à sa disposition. A priori concrète et peu sujette à controverse, cette injonction a mis en lumière la variété et la complexité des interprétations possibles, même parmi ceux exerçant des fonctions et des activités similaires. La définition du niveau de granularité est sans conteste l'un des points de divergence les plus importants. Pour cause, la particularité temporelle et l'arborescence des expériences réalisées à AGLAE, marquées par une forte hétérogénéité et des ruptures de linéarité avec des régularisations correctives fréquentes. Par ailleurs, une même expérience peut être pratiquée sur un seul ou plusieurs objets patrimoniaux, brouillant toute recherche d'unicité par l'intermédiaire de ces derniers. Ainsi, une expérience à AGLAE (*run*), peut être initiée à la suite d'une demande officielle de temps de faisceau ou pas (cas des expériences en interne), dans le cadre d'une demande de service ou d'un projet de recherche, national ou européen (voir PROPOSAL), porter sur un ou plusieurs objets patrimoniaux, s'étendre sur quelques heures à plusieurs jours, et nécessite le plus souvent l'association de plusieurs techniques d'analyse, l'éventail de ces dernières allant de la méthode PIXE (*Particle-Induced X-ray Emission*) à la méthode PIGE (*Particle-Induced Gamma-ray Emission*), en passant par la ionoluminescence et la spectroscopie de rétrodiffusion de Rutherford (RBS).

Par conséquent, pour définir par capillarité le périmètre d'un jeu de données, nous nous sommes d'abord accordés sur le cadre sémantique d'une expérience (*run*). Cette dernière est ainsi d'abord caractérisée par une stabilité du contexte expérimental avec des paramètres et des standards définis en début de manipulation, la moindre modification contrevenant à la cohérence et à la compréhension de l'expérience par ceux qui en sont extérieurs. Une expérience peut être multitechniques, les données PIGE, PIXE et RBS étant générées automatiquement et ne demandant pas de sélection initiale. Elle donne lieu à la création d'un dossier dont le nommage est composé actuellement du nom de l'analyste, de la référence de l'objet et de la date. Un fichier Excel est généré automatiquement

pour toute l'expérience et contenant le numéro de dossier, la référence de l'objet, le nom du projet, la Dose/seconde, la durée de l'analyse, les détecteurs, les filtres, les standards, la taille de la cartographie, l'énergie du faisceau et le type de particule. Dans le contexte d'une expérience réalisée par AGLAE, un jeu de données renvoie donc à l'ensemble des données brutes – fichiers LST et fichiers spectres primaires – et des données extraites via la suite logicielle *ad hoc* pour un *run*, avec le fichier Excel intégrant les métadonnées du contexte expérimental, les données traitées mettant en évidence le phénomène physique voulu, les éléments du cahier de laboratoire, qu'il soit imprimé et/ou dématérialisé, les photographies inhérentes aux zones d'intérêt, les standards, le tout contenu dans un dossier d'utilisateur. À court terme, ce jeu de données devrait correspondre à un fichier en format compressé HDF5, adapté aux données scientifiques massives et hétérogènes, incluant les métadonnées aux niveaux d'arborescence appropriés. Il permettrait le regroupement cohérent et documenté des données issues des analyses ponctuelles et de l'imagerie, en évitant leur fragmentation dispersée.

Après avoir déterminé qu'un jeu de données pour AGLAE équivaut à une expérience (run) définie par une stabilité des paramètres du contexte expérimental, il nous faut aller au plus près des flux de données pour leur faire sens dans une arborescence adaptée et suffisamment souple pour s'adapter à la variété des cas d'usage. De plus, l'hétérogénéité et la discontinuité éventuelle des données imposent le recours à la subsidiarité sémantique, comprise comme l'ajustement des métadonnées au niveau de granularité le plus proche des contenus décrits. Dans un jeu de données inhérent à une expérience, un niveau supérieur

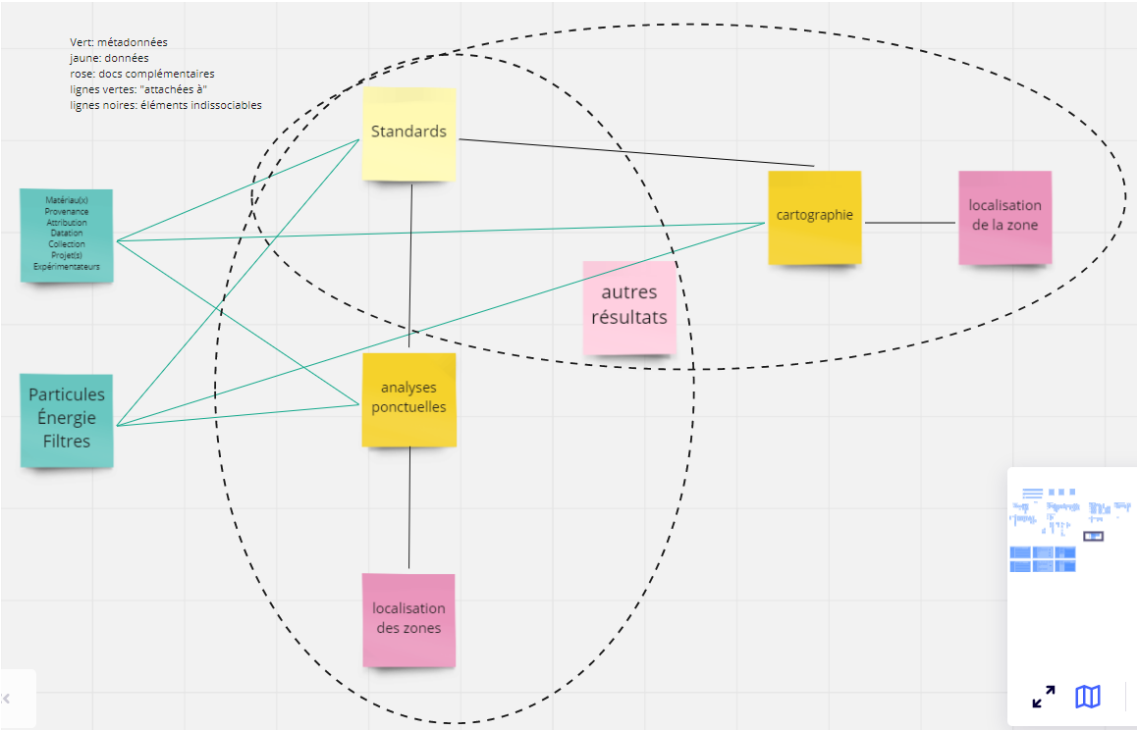


FIGURE 2.1 – Restitution de l'arborescence générale des métadonnées de New AGLAE @New AGLAE

d'arborescence correspondant au contexte expérimental général s'est d'abord imposé de façon évidente, avant de poser la question épineuse de la granularité des informations contenues dans ce champ. Dans un premier temps, en plus de l'énergie du faisceau et du type de particule projetée, nous avons considéré la DOSE par seconde et la taille du pinceau du faisceau comme des attributs du contexte expérimental. Les remarques et corrections apportées par l'équipe d'AGLAE ont permis d'affiner la compréhension des métadonnées et d'en rectifier la portée parfois exagérément étendue ou trop restreinte. Dans le cas présent, la DOSE par seconde et la taille du pinceau du faisceau sont en réalité des attributs affiliés à un « point de mesure », générant la création d'une entité du même nom. Cet ajout a entraîné celui d'une entité « spectre » associé aux entités « point de mesure » et « fichier ». En effet, il est apparu que la première proposition de modèle conceptuel [4], s'arrêtant au contenant et non au contenu des données, souffrait d'un excès de synthétisation, et ne reflétait pas la variabilité scalaire d'une expérience menée à AGLAE. Au cours d'une analyse ponctuelle, les données recueillies correspondent bien à un point précis de l'objet-cible et non à l'intégralité de ce dernier. Ce raffinement conceptuel est absolument déterminant pour la compréhension des choix expérimentaux et de ses résultats. Par ailleurs, c'est justement par la comparaison entre les spectres générés à partir de points de mesure distincts que l'analyste peut confirmer ou infirmer solidement les hypothèses initiales de provenance, de composition et de fabrication. Enfin, ce procédé d'analyse est incontournable dans le cas d'objets composites ou ayant subi des modifications/ajouts au cours du temps. Pour gagner en intelligibilité rétrospective à l'égard du jeu de données, il a donc fallu s'écarter de la zone de confort de l'implémentation et descendre à l'échelle presque microscopique du point de mesure, en augmentant le défi de la masse à structurer et de la scalabilité future. Ce choix de modélisation cristallise la difficulté majeure inhérente à la multiplicité des usages prévus d'Euphrosyne, avec la prise en compte des différentes étapes du cycle de vie des données. En effet, la structure et l'arborescence indispensables à l'utilisateur responsable d'une expérience peuvent être lourdes à implémenter et insatisfaisantes dans une perspective d'accès historique aux données froides, trop massives.

La complexité et l'absence de linéarité du *workflow* concret des utilisateurs ont par ailleurs largement corsé l'effort d'universalité d'application de notre modèle à l'ensemble des cas expérimentaux envisagés. Le traitement conceptuel des standards de l'expérience est en cela un véritable cas d'école qui mérite une attention particulière. Essentiels à la

4. Se reporter au répertoire « MCD_AGLAE_Définitif », lequel contient le modèle conceptuel pour une expérience à AGLAE, ainsi que sa documentation. Un modèle conceptuel est une représentation graphique permettant de « circonscrire un monde » en présentant des entités, leurs attributs et les associations entre les entités avec l'indication de leurs cardinalités (nombre minimum et maximum d'occurrences entre les entités). Un modèle conceptuel prépare le modèle relationnel, indispensable à la réalisation d'une base de données-métier. Le modèle relationnel (Voir le répertoire « MRD_AGLAE_Définitif » contenant le modèle et sa documentation) traduit les cardinalités en clés étrangères et en tables de relation selon que les associations soient simples ou multiples.

fiabilité, à la rigueur et à la compréhension d'une expérience, les standards constituent un maillon particulier et central de la chaîne d'acquisition des données. Dans le monde physique, ils consistent en de modestes pastilles échantillonnées et placées en rang devant la ligne de faisceau. De composition parfaitement connue et référencée, ils sont sélectionnés par l'expérimentateur en fonction du matériau étudié et de la technique d'analyse utilisée. La pertinence de leur utilisation détermine celle de l'expérience elle-même, puisque les données obtenues sont interprétées à l'aide de celles acquises sur les standards. Passés en début et en fin d'expérience, parfois également au milieu de cette dernière, ils participent pleinement de la définition du contexte expérimental. C'est pourquoi, dans un premier temps, une entité « standard » a été créée et associée au contexte expérimental. Les analystes de l'équipe d'AGLAE ont alors soulevé l'impossibilité de limiter un standard à un nom et à une référence alphanumérique, car il acquiert son sens dans un usage dynamique et une savante manipulation qui doit s'inscrire dans un circuit conceptuel. De surcroît, les standards s'inscrivent dans une double temporalité : d'abord celle de leur analyse par la ligne de faisceau, puis celle du traitement croisé des données obtenues avec celles de l'expérience réalisée sur un lot ou un objet-cible. Les standards ont donc une portée à la fois synchronique et diachronique au regard de l'expérience. Ils sont tout autant à sa périphérie que dans son flux interne en générant des fichiers spectres qui serviront à l'interprétation d'autres fichiers spectres. La possibilité de faire figurer un standard en attribut de type booléen d'une entité « spectre » a été évoquée et finalement écartée. En vue de la traduction du modèle conceptuel en modèle relationnel de données, il fallait envisager le respect des formes normales et ne pas définir d'attribut lui-même porteur d'autres attributs qui ne dépendraient pas directement de l'entité. De plus, nous ne souhaitions pas perdre la possibilité de mise en exergue de la spécificité conceptuelle d'un standard, qui ne renvoie pas seulement à un spectre lui-même lié à un point de mesure, mais également à un échantillon doté d'une matérialité à part entière. Par conséquent, nous avons finalement créé une entité « standard » avec ses attributs d'identification matérielle, associé à une entité « point de mesure » reliée à la fois à un spectre et à une technique d'analyse. Adapté à une modélisation respectueuse des flux réels de données, ce choix est bien plus complexe à traduire dans une arborescence de métadonnées, en particulier avec l'incertitude de leur format et du conteneur à venir du jeu de données. Une migration vers le format conteneur HDF5 est toujours en cours d'étude à ce jour.

Déjà utilisé par le synchrotron SOLEIL et d'autres grands instruments d'optique et d'imagerie, HDF5 pour Hierarchical Data Format est un format permettant de sauvegarder et de structurer des fichiers contenant des données massives. Un fichier HDF5 est ainsi un conteneur de fichiers aux formats divers, qui offre déjà une arborescence par sa structure et sa composition en groupes, datasets et l'inclusion d'attributs équivalant à des métadonnées. À l'origine, sa première version exploitable sert aux données de la NASA, qui a besoin d'effectuer des calculs complexes sur de grandes masses de données,

grâce à une indexation efficace. AGLAE ne produit pas de données massives, même dans le cas de l'imagerie, qui ne pèse guère plus que quelques teraoctets. Mais le recours à un format conteneur adapté aux données scientifiques hétérogènes, apte à supporter des fichiers d'imagerie mais également des fichiers d'analyse avec des métadonnées associées aux niveaux d'arborescence adéquats a ouvert une piste de mise en cohérence avec le standard dominant, en vue de l'ouverture et du partage souhaités des données d'AGLAE dans le contexte de l'*Open Science*[5]. Nous sommes donc dans le cas d'un usage du format HDF5 quelque peu détourné de sa vocation initiale et utilisé davantage pour sa capacité à associer le stockage et la description de contenus que pour l'interprétation et le calcul de données massives. Lors de la première phase de création d'Euphrosyne, il est rapidement apparu que si la migration des données d'imagerie telles que celles générées par AGLAEMap vers HDF5 se ferait sans difficulté, il n'en irait pas de même pour les données de l'analyse ponctuelle, dont les formats de fichiers convertis ne sont pour l'heure pas supportés par HDF5. À ce jour, nous ignorons donc encore si ce format sera utilisé pour l'ensemble des données d'AGLAE ou seulement une partie d'entre elles, et s'il sera implémenté dans le cadre de la seconde phase de déploiement – autrement dit la création d'Euphrosyne-data – ou non. Cette incertitude nous a incités à réfléchir simultanément à deux types d'arborescence éventuels : l'un dans le cadre d'une intégration des données dans un format conteneur HDF5 ; l'autre, inspiré de l'organisation actuelle des données et des fichiers qui les renferment.

Dans le premier cas, le principe de subsidiarité sémantique est poussé à son maximum avec la possibilité de renseigner des métadonnées à un niveau de granularité extrêmement fin. Dans un format xml, des informations de contenu et de description seraient associées à chaque fichier ou groupe de fichiers selon les cas, préparant ainsi le travail de pérennisation et d'archivage. L'arborescence serait fragmentée et ne serait pas aisément compréhensible dans une perspective d'ensemble, mais elle aurait le mérite d'être conçue sur mesure et de s'adapter à la forte plasticité des données d'AGLAE. Dans le second cas, un document xml valable pour l'intégralité d'un jeu de données serait produit et renseigné au moyen d'une extraction automatique des métadonnées incluses dans le corps des fichiers, et de leur attribution dans des champs préalablement balisés. Ce procédé impliquerait certes, une certaine rigidité structurelle et sémantique, mais aussi une cohérence harmonieuse pour tous les jeux de données qui y seraient soumis. Un niveau d'arborescence général relatif au contexte expérimental serait lui-même subdivisé par techniques d'analyse (PIXE, PIGE, RBS, IBIL, etc.), découpées en types d'analyse (imagerie, analyse ponctuelle) contenant eux-mêmes un niveau inférieur correspondant à chaque point de mesure. Dans le cas des métadonnées relatives aux données issues de l'association de plusieurs techniques d'analyse – ce qui représente une large proportion des expériences - , elles bénéficieraient d'un niveau

5. Présentation d'audience d'Euphrosyne pour l'obtention du statut de start-up d'État du Ministère de la Culture, 8 avril 2021.

d'arborescence à part entière, contenu dans celui du contexte expérimental, et profiteraient de l'utilisation de pointeurs vers les techniques et les fichiers initiaux concernés. Nous nous

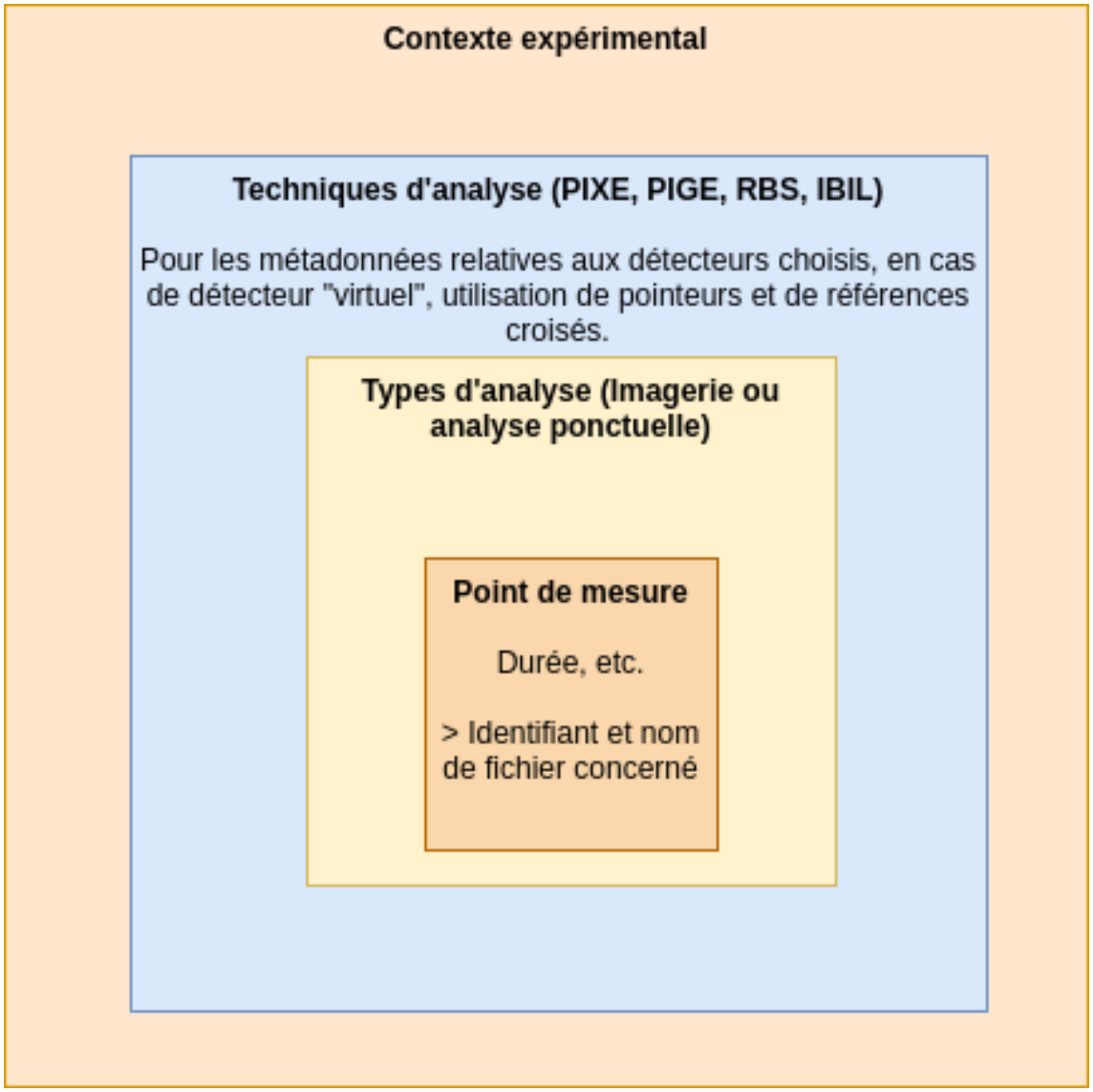


FIGURE 2.2 – Schéma récapitulant l'arborescence générale des métadonnées pour une structuration en xml.

sommes inscrits dans la perspective d'un utilisateur ou d'un individu extérieur susceptible de vouloir accéder uniquement aux données traitées définitives ou d'exploiter et croiser des données primaires et intermédiaires peu ou pas utilisées lors de l'expérience. À l'instar de la description archivistique numérique, les métadonnées seraient renseignées au niveau le plus approprié et ne seraient donc pas répétées aux éventuels niveaux inférieurs.

2.1.3 Euphrosyne : l'accès aux données pour épicentre

Qu'il s'agisse d' Euphrosyne-manip ou d'Euphrosyne-data, correspondant respectivement à la première et à la seconde phase de déploiement de l'interface, la création

d'un accès optimisé et inédit aux données conforme à la déontologie FAIR constitue le centre névralgique du projet. Euphrosyne-manip permettra un premier niveau de décloisonnement en autorisant la visualisation et le traitement de s données expérimentales à distance, avec un renseignement à la fois manuel et automatique des métadonnées en adéquation avec la nouvelle arborescence et les champs de référence définis conjointement avec les utilisateurs. Le format xml des métadonnées favorisera les échanges entre applications et systèmes ainsi que leur interrogation au-delà de l'interface. Euphrosyne-data amorcera un niveau de progression supplémentaire en offrant une fenêtre et des outils d'accès aux données définitives. Afin d'assurer leur sécurité et leur intégrité, il serait nécessaire d'opter pour un système de stockage et de conservation externalisé, telle qu'une infrastructure spécialisée dans l'hébergement et la valorisation des données scientifiques dans les différentes étapes de leur cycle de vie. Les partenariats noués récemment entre le Cines, Huma-Num et d'autres infrastructures érigées sur le principe de pérennisation et d'accès aux données de la recherche, laissent entrevoir la possibilité d'une solution « toute-en-un » où les besoins de stockage, d'accès, d'agrégation sémantique et d'archivage des données seraient pris en charge dans la pluralité de leur cycle de vie au sein d'une même entité, en fonction d'une plannification élaborée en étroite association avec les spécialistes de la médiation et de l'archivage numériques. Le recours à une solution externalisée s'accompagnerait du maintien d'une procédure interne de stockage à AGLAE, où les données d' Euphrosyne seront répliquées. Ainsi, l'exigence d'accès et d'ouverture universels aux données sera conciliée à celle des besoins immédiats des opérateurs d' AGLAE et des utilisateurs. Grâce à une politique de gestion d'identifiants telle que définie précédemment, ainsi que par une interopérabilité à la fois technologique et sémantique, les données d'Euphrosyne pourront également être interrogées au moyen des protocoles http, facilitant leur découverte et leur récupération.

2.2 *Interoperable, reusable* : favoriser l'interopérabilité et la réutilisation des données

L'interopérabilité, qui doit permettre à des programmes/machines/applications divers de communiquer ensemble sera à la fois interne et externe. Si la première forme sera garantie par l'harmonisation du format et de la structure des métadonnées nécessaires à l'interrogation et à l'agrégation des données générées par les différentes techniques d'analyse du C2RMF, la seconde, dans ses dimensions sémantiques et syntaxiques reposera avant tout sur l'élaboration de référentiels et de normes conçus actuellement dans les groupes de réflexion et de travail DIGILAB [6]. En cela, la figure de proue qu'est la

6. Organe d'E-RIHS, infrastructure d'excellence et d'intérêt global pour l'étude des matériaux du patrimoine, DIGILAB doit permettre l'accès à une infrastructure numérique pour le traitement des données quantitatives produites par les différents instruments, selon une politique respectueuse des principes

construction de l'ossature conceptuelle du catalogue des données d'Euphrosyne joue un rôle pivot.

En tant qu'opération intellectuelle d'identification et de signalement, le catalogage des données d' AGLAE devra s'appuyer sur un ensemble de références conçues et validées transversalement, à la fois par la communauté-métier spécialiste des contenus et par les spécialistes des systèmes d'information et de documentation. On parle alors de référentiel, terme générique recouvrant une réalité protéiforme, de la simple liste de mots au thésaurus. Il s'agit dans un premier temps de contrôler les formes des termes utilisés lors du catalogage afin d'éviter les redondances d'information et d'offrir des clés permettant la classification ainsi que la recherche aisée des contenus [7]. Le référentiel peut notamment prendre la forme d'un thesaurus, vocabulaire contrôlé hiérarchique le plus fréquent. Avant d'opter pour un référentiel spécifique, il est d'abord question d'une entente sémiologique et sémantique à créer à la fois au sein de la communauté élargie d' AGLAE – composée à la fois des spécialistes en sciences des matériaux et des acteurs des institutions patrimoniales - , et parmi la communauté plus étendue de l'analyse par faisceau d'ions. Selon une dynamique centrifuge, il a semblé naturel et opportun de s'appuyer d'abord sur le premier cercle formé par la communauté d' AGLAE sans toutefois l'émanciper complètement de la sphère de l'analyse par faisceau d'ions à laquelle elle appartient. Précisons d'emblée qu'en dépit des attentes explicitement et récemment formulées d'une partie des membres de cette sphère en la matière [8], il semble intellectuellement impossible de couvrir en un seul référentiel l'ensemble des activités et des résultats extrêmement divers produits par l'analyse par faisceau d'ions. La classification réalisée pour les données d'AGLAE est donc vouée à être un sous-ensemble d'une classification mosaïque plus vaste qui reste aujourd'hui à construire.

Notons que la spécificité de la terminologie d' AGLAE est telle qu'elle offre au moins l'avantage de ne guère offrir la possibilité de synonymes ou de fluctuation lexicale. Les variations concernent ainsi plutôt les objets de l'expérience. Chaque élément de cette terminologie nécessite d'être décomposé conceptuellement en vue d'une représentation logique adaptée à la granularité du signalement des (méta) données. L'émulation collective et les circulations conceptuelles diffusés à travers les canaux de DIGILAB, d'ESPADON et de SSHOC (*Social Sciences and Humanities Open Cloud*) [9] ont permis d'ajuster et

FAIR.

7. Maxime Challon, Les référentiels en institutions patrimoniales : évolution des pratiques et repositionnement. L'exemple des référentiels de l'Institut National de l'Audiovisuel, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Gautier Poupeau, École nationale des Chartes, 2020, p. 17.

8. Échanges menés dans le cadre de la 8e Rencontre « Ion Beam Applications Francophone », organisée du 5 au 7 juillet 2021 par la Société Française du Vide.

9. *Social Sciences Humanities Open Cloud* (SSHOC) est un projet financé par le programme-cadre de l'Union Européenne Horizon 2020. Il réunit une vingtaine d'organisations partenaires – notamment des organismes patrimoniaux - afin de développer un *cloud* ouvert et accessible dédié aux ressources en sciences humaines et sociales. Les partenaires possèdent tous une expertise sur l'ensemble du cycle de vie

d'alimenter notre essai de classification en nous inspirant des plus récents travaux en matière de référentiel et d'organisation sémantique des données des sciences du patrimoine.

Parce qu'il entremêle les enjeux de pérennisation, d'interopérabilité, de recherche scientifique et de conservation patrimoniale, l'exemple de la National Gallery (NG) [10] nous a inspiré le choix d'un certain nombre d' objets, ressources et prédicats pour notre propre modèle. Il s'agit pour AGLAE de pouvoir partager ses jeux de données expérimentaux tout en les reliant prioritairement :

1. À des ressources internes relatives aux objets patrimoniaux analysés et aux résultats des autres techniques d'analyse du C2RMF, actuellement répartis entre la base EROS et les différents silos du laboratoires.
2. À des ressources mutualisées dans le cadre d'ESPADON [11], de DIGILAB et plus largement d'E-RIHS.
3. À des bases de données dédiées au patrimoine culturel telles que Joconde.
4. À des bases de données-métiers consacrées à l'analyse par faisceau d'ions et aux grands instruments de physique des particules.
5. Aux ressources et outils des technologies IIIF.

Cette feuille de route nous a incité à raffiner notre modélisation selon une triple approche, relativement similaire à celle de la NG : la représentation contextualisée de l'objet patrimonial à l'aide des ressources du CIDOC-Crm [12], largement utilisées et déclinées par les différentes structures de conservation et de médiation patrimoniales françaises et étrangères ; celle des procédures expérimentales appliquées au patrimoine à partir de certains éléments du CIDOC-Crm Sci [13] mais également de l'ontologie appliquée à la biochimie [14] ; celle des objets numériques inhérents aux résultats de ces procédures.

---

des données, de leur production à leur conservation en passant par leur réutilisation et leur pérennisation. Site web : `https://sshopencloud.eu/`

10. Présentation des exemples inspirés de CIDOC-Crm, CIDOC-Crm Sci et de CIDOC-Dig pour la modélisation conceptuelle des ressources inhérentes aux recherches menées sur l'œuvre de Raphaël [Site web de la National Gallery : `https://jpadfield.github.io/sshoc-ng/Raphael%20Examples.html`. Consulté le 15 juin 2021.]

11. La dénomination recouverte par ce dernier sigle mérite une attention particulière : « En Sciences du Patrimoine, l'Analyse Dynamique des Objets anciens et Numériques ». Sous couvert d'une mise en relief de l'événement que constitue la technique expérimentale appliquée au patrimoine, l'ambition de ce projet d'Équipement d'Excellence (EquipEx) est en réalité moins technologique que documentaire et sémiologique. Porté par la Fondation des Sciences du patrimoine et sélectionné par le Programme d'Investissements d'Avenir (PIA 3), ESPADON doit créer « l'objet patrimonial augmenté » selon une double sémantique : scientifique d'une part, avec le perfectionnement des techniques d'analyse par l'amélioration des instruments existants ; numérique d'autre part, avec la mutualisation et l'enrichissement des données produites.

12. CIDOC-CRM, url : http ://www.cidoc-crm.org/ [Consulté le 16 mai 2021]. La première version du CIDOC-Crm est achevée en 1999. En 2006, ce modèle a fait l'objet d'une publication ISO, en devenant une norme internationale : ISO 21127 : 2006.

13. Url : `http://www.cidoc-crm.org/crmsci/home-1` (Consulté le 07 juillet 2021).

14. BioPortal, portail consacré aux ontologies appliquées à certains sous-domaine de la biologie. Url : `https://bioportal.bioontology.org/ontologies/CHMO/?p=classes&conceptid=root`

Entre la proposition du CIDOC-Crm Cr et la piste hybride ouverte par la NG, nous avons ainsi esquissé une troisième voie, adaptée à la singularité des données d' AGLAE ainsi qu'aux usages actuels et futurs prévus pour elles. Notre principal apport réside dans le niveau de granularité de l'expérience, décomposée jusqu'à atteindre son objet à l'échelle de l'infiniment petit : l'élément chimique. Nous nous sommes efforcés de représenter ce processus dans le respect de sa double dynamique, à savoir le permanent et l'occurrent [15], en lui conférant la souplesse conceptuelle nécessaire pour s'adapter à l'imprévisibilité et à la discontinuité temporelle. L'ontologie appliquée à la biochimie a pour cela été particulièrement utile à notre réflexion, puisqu'elle est organisée selon deux axes principaux : « continuant », qui regroupe l'entité matérielle analysée et ses dépendances ; et « process » qui comprend les événements liés à l'expérience tels que les accidents de laboratoire, la plannification du processus expérimental et ses éventuels réajustements. Nous nous en sommes inspirés pour la modélisation générique d'un contexte expérimental à AGLAE avec la création d'un triplet, directement relié par sa ressource à celui consacré à la définition de l'événement général (crm : E5 Event/ is defined/ by BCO : experimental_planning_process) :

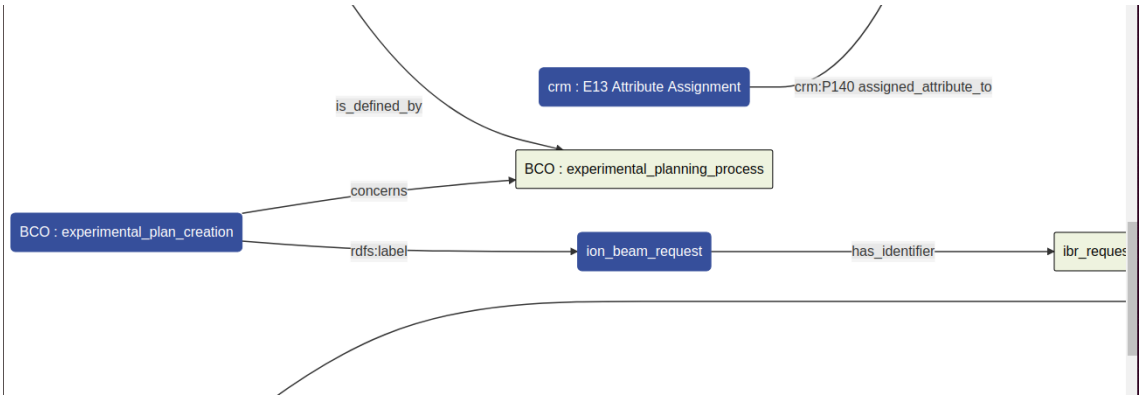


FIGURE 2.3 – Capture d'écran d'un extrait du modèle sémantique élaboré pour une expérience AGLAE.

La difficulté majeure de notre conceptualisation a résidé dans notre volonté de restitution étroite de l'interdépendance des ressources numériques (crm : dig) avec les flux de l'expérience elle-même et la matérialité complexe et protéiforme des objets patrimoniaux étudiés. En effet, les données d' AGLAE trouvent leur substance cognitive dans cette jonction tant technologique que scientifique et culturelle. La structure en graphe en a facilité la visualisation et la répartition interconnectée des différents objets qui en découlent. Toutefois, notre modèle aurait dû traiter avec davantage de finesse les ressources relatives à certains paramètres de l'analyse telles que la DOSE ou le facteur de calibrage du spectre. Le chevauchement sémantique de ces éléments, qui relèvent à la fois de l'expérience et de l'interprétation de ses résultats, ainsi qu'une maîtrise relativement superficielle de leur

15. Gilles Kassel, « Une alternative à la distinction 'continuant' vs 'occurrent' », dans 29e Journées Francophones d'Ingénierie des connaissances, Nancy, 2018, p. 147-162.

contenu scientifique – particulièrement complexe pour un individu non-initié – ont justifié cette prudence temporaire, dans l'attente de développements futurs en concertation avec des spécialistes aguerris. L'utilisation d'identifiants ARK rendra possible l'agrégation des ressources d' AGLAE avec des contenus externes.

La prochaine étape sera la présentation de notre référentiel et de sa représentation conceptuelle auprès des institutions partenaires de DIGILAB et d'ESPADON, ces programme et équipement d'excellence étant actuellement à l'œuvre pour l'édification d'un socle sémiologique commun. Une phase déterminante d'itération collective s'ensuivra jusqu'à atteindre un modèle consensuel. Un vaste chantier devra également être entrepris pour essaimer au moins partiellement notre proposition de classification auprès de la communauté élargie de l'analyse par faisceau d'ions, dont les ressources sont pour la plupart cloisonnées dans des silos applicatifs.

Table des matières