Huma-Num IR*

distam
Digital Studies
Africa
Asia
Middle East

Le signalement numérique des ressources documentaires au format EAD

Un état des lieux des logiciels libres – Décembre 2022

Rapport rédigé par Mehdi Sakatni dans le cadre d'une mission financée par le GIS Moyen-Orient et mondes musulmans (plan SHS) pour le consortium DISTAM en 2021-2022

Sommaire

1. INTRODUCTION

Ce guide est consacré aux logiciels libres (ou Open Source Softwares) pour le catalogage et la description de manuscrits et d'archives au format EAD (Encoded Archival Description).

Les premiers logiciels libres de description archivistique sont nés au début des années 2000. Ils s'inscrivaient alors au croisement de deux dynamiques :

- la promotion du principe de l'*open source* dans le développement des outils informatiques (naissance de l'Open Source Initiative en 1998)
- la standardisation des langages de description archivistique avec l'adoption du format EAD (Encoded Archival Description) en 2002 qui permet une description multi-niveaux et facilite la recherche en affichant la structure hiérarchique des fonds

Les nouvelles normes de métadonnées offraient la possibilité d'accroître l'accès aux informations sur les documents d'archives et de promouvoir de nouveaux processus archivistiques. Afin d'exploiter ces possibilités, les archivistes avaient besoin d'un nouvel outil leur permettant la description et la gestion des fonds. La mise en œuvre du format EAD repose sur l'exploitation du langage XML (Extensible Markup Language). Ce format permet l'interopérabilité mais il est soumis à des contraintes de règles de fonctionnement du langage qui impliquent le recours à des éditeurs spécifiques et limitent le choix de logiciels envisageables pour produire de l'EAD.

C'est pour répondre à cette demande que de premiers logiciels libres ont été développés dans la première décennie du XXIe siècle, quasi-exclusivement en Amérique du Nord.

1.1 Périmètre et enjeux de l'étude

- Une solution logicielle libre et pérenne

Le premier enjeu consiste à s'ancrer dans le domaine de la science ouverte. Il s'agit donc de privilégier une solution libre de droits, fondée par ou appuyée sur une communauté d'utilisateurs suffisamment solide et active pour en garantir la pérennité sur le long terme.

Les logiciels analysés dans le cadre de cette étude ont par conséquent publié leur code source en ligne, sous licence ouverte qui permet à chacun de consulter, d'utiliser et de modifier le code pour l'améliorer, à la condition de le partager.

Les récentes études du paysage des logiciels de description documentaire montrent que le modèle du logiciel propriétaire reste dominant[1]. Dans un tel contexte, le choix d'un logiciel libre est tout sauf une évidence. Pourtant, le critère premier de sélection d'un logiciel libre, selon l'enquête réalisée par Lyrasis auprès des utilisateurs en 2021, est bien celui de la durabilité, avant celui du coût financier[2].

- Interopérabilité

Le deuxième critère de sélection concerne l'adéquation des formats de description et de production des métadonnées aux normes internationales en vigueur. L'adéquation aux normes garantit l'interopérabilité du logiciel (et donc des descriptions produites via celui-ci) avec les différents outils institutionnels (le logiciel Calames utilisé par l'ABES par exemple, articulé au référentiel d'autorités Idref ; le logiciel TAPIR utilisé par la BnF) et, par là-même, le

1 https://www.nationalarchives.gov.uk/archives-sector/advice-and-guidance/managing-your-collection/documenting-collections/cataloguing-and-archives-networks/.

2 Hanna Rosen, Jill Grogg, *LYRASIS 2021 Open Source Software Survey Report. Understanding the Landscape of Open Source Software Support for American Libraries*, 2021, p. 5.

versement des instruments de recherche vers des entrepôts institutionnels assurant la pérennité, l'interopérabilité et la réutilisation des contenus.

- Multilinguisme

Une des activités spécifiques aux équipes du consortium DISTAM consiste en la production de notices de description en écritures non-latines. Les logiciels de description archivistique visés sont donc des programmes en mesure de produire et de supporter des contenus multilingues, et, en particulier, de combiner des écritures dextroverses et sinistroverses dans une même notice.

- Accessibilité

L'existence de logiciels libres de droits permet aux utilisateurs de petites institutions (bibliothèques privées, associations, sociétés savantes, …), pas forcément dotées d'outils adaptés, de se saisir de ces logiciels et de produire des descriptions de contenus. Dans le cadre de la recherche universitaire, les logiciels de description peuvent également être mis à profit par des chercheurs, afin de créer de nouvelles notices ou d'enrichir des notices existantes. Cela n'est réalisable que si le logiciel permet un environnement ergonomique d'édition du xml, par exemple en mode formulaire. On privilégiera par conséquent des programmes informatiques présentant une interface facilement manipulable et éventuellement paramétrable selon les besoins spécifiques des utilisateurs.

L'accessibilité inclut aussi le fait que l'installation de l'instance logicielle soit relativement simple et pratique à réaliser sur des postes de travail informatique dotés d'une interface classique, sans dépendance avec un outil propriétaire ou un OS propriétaire.

L'objet de ce guide est d'explorer les solutions logicielles libres pour la description et le signalement des ressources documentaires. Par l'analyse de l'environnement institutionnel, de la structuration technique et de la gamme des services accessibles, nous proposons de faire un état des lieux des logiciels disponibles à ce jour et de formuler des recommandations aux usagers. Ce guide, du fait de son format et son périmètre, ne peut prétendre à l'exhaustivité ; dans un champ qui s'est considérablement développé au cours de la dernière décennie, les logiciels de description documentaire se sont multipliés[3]. Au sein de cette offre déjà conséquente, nous avons sélectionné un échantillon de quatre logiciels libres, afin d'en étudier les forces et les faiblesses et d'en faire ressortir les spécificités. Les quatre logiciels qui font l'objet de cette étude sont :

- Access to Memory (AtoM) : https://www.accesstomemory.org
- ArchivesSpace : https://archivesspace.org/
- CollectiveAccess : https://www.collectiveaccess.org/
- Hyku : https://hyku.samvera.org/

Ce guide s'appuie en partie sur une étude du logiciel AtoM réalisée en 2022 par les entreprises Six&Dix et l'Informatique Communicante (LIC), et financée par le GIS Moyen-Orient et Mondes Musulmans.

1.2 Présentation générale des solutions logicielles étudiées

Access to Memory (AtoM)
AtoM est un logiciel libre pour la description archivistique et l'accès aux instruments de recherche sur le web. Il a été développé depuis 2007 par la société canadienne Artefactual Inc. avec le soutien, à l'origine, de l'ICA (International

3 Les Archives nationales du Royaume-Uni, dans un inventaire non exhaustif des solutions logicielles en 2020, ont recensé 38 outils (licences libres et logiciels propriétaires confondus) pour la gestion numérique des collections : https://www.nationalarchives.gov.uk/archives-sector/advice-and-guidance/managing-your-collection/documenting-collections/cataloguing-and-archives-networks/

Council on Archives). AtoM produit des notices de description multilingue au format EAD. Le logiciel s'appuie sur une communauté internationale importante.

ArchivesSpace

Le projet ArchivesSpace a été lancé en juin 2009, autour de la fusion des applications Archivists' Toolkit et Archon, pour une sortie de la V1 en septembre 2013. ArchivesSpace est développé par Lyrasis et s'appuie sur une communauté de plus de 450 utilisateurs dans le monde, dans laquelle on trouve en majorité des institutions universitaires nord-américaines, mais également des bibliothèques publiques, des musées et des archives d'associations. ArchivesSpace peut gérer des contenus analogiques, numériques et hybrides.

CollectiveAccess

CollectiveAccess se définit comme un logiciel en accès ouvert pour le catalogage et la publication de collections muséales et archivistiques. Il a été développé par la société américaine Whirl-i-Gig à partir de 2003, en collaboration avec des institutions partenaires en Europe et en Amérique du Nord. CollectiveAccess est l'outil privilégié par plusieurs centaines d'institutions présentes sur les cinq continents. Les centres d'arts et les musées sont les institutions les plus représentées parmi les utilisateurs de CollectiveAccess.

La société Whirl-i-Gig propose un ensemble de services autour de son logiciel (accompagnement, projet de conservation numérique, développement de sites web, migration de données) qui permettent de définir les besoins des utilisateurs afin d'améliorer le produit. La dernière version 1.7.14 a été publiée en janvier 2022.

Hyku

Hyku se définit comme la « solution de dépôt numérique nouvelle génération » ; le logiciel a été créé en réponse à l'appel de l'IMLS (Institue of Museum and Library Services) pour la mise en place d'une plateforme numérique nationale américaine.

Hyku présente à la fois les fonctions de CMS (Collection Management Service) et de DAMS (Digital Asset Management Service). L'ambition est de dépasser les caractéristiques des précédentes solutions logicielles, limitées car pour la plupart créées lors d'une autre époque du web. Les logiciels précédents avaient pour objectif premier la mise en ligne des collections. Aujourd'hui, les contenus numériques, les flux de travail qui leur sont appliqués et les mécanismes de publication de contenus sur le Web sont tous devenus beaucoup plus sophistiqués. Hyku a pour ambition de répondre à cette nouvelle donne.

La version 1.0 a vu le jour en 2019.

2. ENVIRONNEMENT DES LOGICIELS

2.1 Les communautés

Le rôle des communautés est essentiel car c'est de leur activité qu'émergent les besoins et les requêtes qui feront évoluer les solutions logicielles. Il faut à la fois des communautés actives (qui font un usage régulier du logiciel pour en cerner les points forts à consolider et les faiblesses à améliorer) et bien organisées (afin que les besoins qui en émergent puissent parvenir jusqu'aux développeurs et être traduites en améliorations concrètes).

Les communautés des logiciels identifiés sont dans l'ensemble vastes et actives, à l'exception de celle de Hyku (dont la faiblesse peut s'expliquer par le caractère récent du logiciel). Elles comptent souvent plus d'une centaine d'institutions utilisatrices de leurs services (plusieurs centaines pour CollectiveAccess à près de 500 pour

ArchivesSpace). Cette forte présence dénote à la fois une demande importante des utilisateurs en matière de CMS et une grande variété dans les besoins.

2.1.1 Les utilisateurs

Les communautés d'utilisateurs sont le reflet des usages et des orientations propres à chaque logiciel étudié. On note ainsi une présence accrue des institutions muséales au sein de la communauté CollectiveAccess (plus orientée description d'objets et de collections des arts visuels), une très forte majorité d'établissements universitaires chez ArchivesSpace (spécialisé dans le traitement des collections des bibliothèques et archives pour la recherche universitaire).

AtoM dispose d'une communauté large et internationale, parmi laquelle on compte des organismes significatifs. En France, AtoM est utilisé par le projet OpenJerusalem, par l'École française d'Extrême-Orient, par le Musée d'Archéologie nationale, ou encore par l'UNESCO.

ArchivesSpace (du fait du système d'*adhésion*) a fait le choix d'une communauté à deux vitesses : d'une part, une communauté d'utilisateurs à proprement parler (qui emploient le logiciel) et communiquent parfois sur le forum d'entraide ; d'autre part, les membres d'AS qui font partie des structures de l'organisme, participent aux comités et ont régulièrement des réunions pour échanger sur les difficultés et les besoins des utilisateurs. ArchivesSpace s'appuie sur un réseau d'institutions universitaires de premier plan, essentiellement localisées aux États-Unis, ainsi que sur une administration du logiciel extrêmement structurée, avec des organismes aux rôles clairement identifiés et fortement hiérarchisés.

2.1.2 Les développeurs

Le développement est souvent le parent pauvre dans la structure institutionnelle des logiciels libres. Dans la plupart des cas étudiés, les ressources humaines en la matière sont extrêmement limitées.

Pour AtoM, par exemple, le maintien et les mises à jour logicielles sont assurés par une petite équipe de développement, avec une personne-clé en l'occurrence Dan Gillean, d'Artefactual. ArchivesSpace, développé par Lyrasis, bénéficie des ressources humaines mutualisées de l'organisme. Un gestionnaire de programmation est ainsi détaché par Lyrasis pour les besoins spécifiques d'ArchivesSpace, mais le Program Manager en question ne se consacre probablement pas à AS à temps complet. De manière générale, les ressources humaines allouées par les institutions-membres en tant que contribution au développement et à l'administration des logiciels libres sont faibles, mais elles sont particulièrement lacunaires dans le domaine technique. Les institutions manquant généralement de personnel qualifié, il leur est difficile d'investir des ressources humaines dans le développement des logiciels, encore moins sur le long terme. L'investissement en ressources humaines sur les aspects autres que techniques (administration fonctionnelle, animation de la communauté, formation, documentation) est plus fréquent, sans être systématique[4].

Au vu de la faiblesse des ressources humaines sur le plan technique, un aspect crucial pour la pérennité des logiciels réside dans la relation entre la communauté d'utilisateurs et l'entreprise informatique qui assure le développement et la maintenance. Dans l'écosystème des bibliothèques universitaires américaines, la société Lyrasis joue un rôle-pivot. Elle est associée, en tant que créateur des programmes informatiques, développeur, ou simplement comme membre des communautés, à toute une gamme de services complémentaires pour les métiers de l'information scientifique et technique.

Lyrasis se présente comme un organisme à but non lucratif dédié à l'accessibilité de l'héritage scientifique et culturel mondial.

Lyrasis a développé ou développe un ensemble de logiciels à destination des professionnels de l'information scientifique et technique (CMS, DAMS, entrepôts numériques). Le rôle de l'organisme est d'identifier les besoins, les difficultés et les défis de ses membres pour leur proposer des produits et des services adaptés.

Lyrasis a aussi développé une réflexion poussée sur les enjeux les plus saillants autour du logiciel libre. Un rapport de 2021 (https://research.lyrasis.org/items/554e7dee-c830-4e52-a7f1-621fa446bc79), synthèse d'une enquête

4 Hanna Rosen, Jill Grogg, *LYRASIS 2021*, *op.cit.*, p. 14.

auprès des usagers, présente leurs recommandations pour parvenir à la réalisation et au maintien de produits qui, à la fois, respecteraient les principes de la science ouverte et répondraient aux exigences pointues d'une sphère économique concurrentielle et en évolution permanente.

Parmi les services développés par Lyrasis :

- ArchivesSpace
- DSpace
- Islandora
- Fedora
- VIVO
- CollectionSpace

2.2 Structuration et financement

Logiciel libre ne signifie pas gratuité. Le choix d'un logiciel libre inclut forcément des coûts de maintenance et de développement que les utilisateurs doivent prendre en compte au moment d'investir l'outil afin de déterminer un modèle économique viable et pérenne.

Il se dégage deux types de structuration et de financement autour des solutions logicielles étudiées.

1. Le logiciel est créé et développé par une entreprise de développement informatique. Elle en propose une version libre autour de laquelle se construit une communauté d'utilisateurs. Ces utilisateurs sollicitent l'entreprise pour des services payants liés au logiciel (développements spécifiques, extensions, création d'un site web utilisateur…). Les améliorations éventuelles produites au cours de ces missions de services sont par la suite intégrées à la version libre du logiciel.

   Ex : Artefactual pour AtoM, Whirl-i-Gig pour CollectiveAccess.

La faiblesse de ce type de structuration est qu'elle est très dépendante des intérêts économiques de l'entreprise-mère. Si cette dernière se désengage du logiciel, la communauté se voit dans l'obligation de trouver une autre entreprise de développement capable de fournir le même type de services, ou bien de changer de modèle de financement.

C'est la situation qui s'est produite avec AtoM et Artefactual. Le désengagement du principal développeur limite fortement la gamme de services proposés et jette le doute sur le potentiel d'évolution du logiciel. Dans l'un de ses messages à la communauté AtoM, Dan Gillean offre un « état de l'art » d'AtoM [10] : Dan Gillean explique que le cœur d'AtoM s'appuie sur une base de Symfony (framework PHP qui est aujourd'hui dépassé) et qu'il devient urgent de revoir l'écriture de l'application. Cependant, il constate que cela demande énormément de temps, de recherche, de développement et d'argent. Réécrire AtoM dans un nouveau code de base est un plan à long terme pour transformer les modules cœurs en un ensemble beaucoup plus modulaire. À cause de cela, il se trouve obligé de limiter l'addition de nouvelles fonctionnalités pour maintenir la stratégie de développement qui devra évoluer éventuellement vers un nouveau code.

Pour parer à ces difficultés, la communauté d'AtoM a constitué une fondation en 2018 (https://accesstomemoryfoundation.org/). Elle se donne pour mission la mise en place d'un nouveau modèle de financement qui permettra la refonte du logiciel et sa mise à jour vers les standards informatiques. Cette nouvelle structure de financement devant mener à AtoM 3 n'est pour l'instant qu'à l'état d'ébauche mais elle semble prendre la direction d'une communauté de membres qui couvriraient les coûts de développement du logiciel par le biais d'une adhésion, se rapprochant par là-même du second modèle identifié.

2. Le logiciel est soutenu par une communauté de membres (institutions) versant une adhésion annuelle (variable selon les besoins ou la taille de l'institution) qui sert à couvrir les frais de fonctionnement de l'équipe de développement.

Ex : ArchivesSpace, Hyku (via la communauté Samvera)

Ce type de structuration apparaît plus solide sur le long terme, mais il est délicat à mettre en œuvre, surtout dans les premiers temps : si la communauté n'est pas suffisamment étendue, la viabilité du projet est dépendante d'autres sources de financement. De plus, il pose question quant au caractère « ouvert » du modèle : si le principe de l'ouverture du code source est toujours respecté, il n'en est pas de même pour les décisions sur les futurs développements, sur lesquels les utilisateurs non-membres n'ont pas de visibilité ni d'influence.

La question centrale est celle de l'évaluation des coûts d'un logiciel libre. Les communautés des logiciels libres sont confrontées depuis leurs débuts à cette question de la pérennité financière et de la maintenance de l'outil. Il n'existe probablement pas de solution miracle mais une réflexion en amont sur les moyens de pérenniser économiquement un logiciel libre semble indispensable pour éviter le déclin et l'abandon des solutions logicielles. Le modèle de l'adhésion payante est à ce jour le modèle privilégié ; en contrepartie, il crée des communautés à deux vitesses, divisées entre membres, payant leur adhésion et ayant un pouvoir décisionnel dans la gouvernance du logiciel, et les utilisateurs, dont l'impact sur l'évolution du produit est limité.

3. ASPECTS TECHNIQUES

Les logiciels étudiés s'appuient tous sur des applications existantes, soit parce qu'il s'agit de refontes de logiciels antérieurs (comme ArchivesSpace qui a repris Archivist's Toolkit et Archon), soit parce que des briques logicielles solides existent pour gérer les contenus web (Drupal) ou pour répondre spécifiquement aux besoins de la description documentaire (Fedora).

L'enjeu est ici de savoir quels sont les environnements de déploiement les plus classiques et les plus sûrs (qui demanderont le moins de maintien).

La structure abrégée en LAMP (acronyme pour Linux Apache MySQL PHP) est la structure la plus employée par les logiciels libres. Ce pack logiciel comprend un système d'exploitation, un serveur *http*, un système de gestion de base de données et un langage de programmation interprété qui permet de gérer du contenu web. Étant la structure logicielle privilégiée par les logiciels libres, elle est régulièrement entretenue et améliorée. Elle est donc relativement facile à déployer et à maintenir. C'est l'architecture employée par CollectiveAccess et AtoM.

En dépit de l'emploi de cette structure courante, AtoM souffre d'un retard considérable dans les mises à jour des briques logicielles. L'installation d'AtoM avec les dernières versions de ces briques est difficile.

De plus, sa dépendance vis-à-vis d'autres logiciels le rend plus vulnérable, notamment lorsque les versions logicielles ne sont pas à jour. C'est le cas, par exemple, de la version d'Elasticsearch utilisée par AtoM : Elasticsearch 5.6 utilise un fichier Java cœur, « log4j » en version 2.11. Le 10 décembre dernier, une vulnérabilité conséquente a été trouvée sur les versions antérieures à 2.17 de ce fichier Java, permettant l'intrusion dans le système avec de grandes conséquences.

L'équipe d'AtoM a été avertie du problème sur le Google Groups par des utilisateurs et a trouvé une solution temporaire au problème sur un AtoM installé en version 2.6. Mais il en est de la charge des administrateurs et des installateurs de l'outil de connaître la vulnérabilité et appliquer la correction (il n'existe pas d'alerte dans la documentation d'installation pointant cette vulnérabilité).

3.1 Formats d'import-export et métadonnées

Le format de production de métadonnées privilégié est l'EAD (Encoded Archival Description).

Il est souvent complété par d'autres formats courants, type ISAD(G), XML ou DublinCore.

- Fonctions d'import

Les formats d'import les plus courants (CSV, XML, XML-EAD, MARC) sont disponibles pour toutes les applications étudiées. Certains logiciels proposent une gamme très large de formats d'import, comme CollectiveAccess et Hyku qui couvrent à peu près tous les formats de fichier en import pour les objets.

Un des défauts d'AtoM, relevé à l'usage par l'EFEO, réside dans la gestion des niveaux de description dans les imports en EAD.

- Fonctions d'export

AtoM, ArchivesSpace, CollectiveAccess et Hyku proposent les formats CSV et XML (EAD) en export.

CollectiveAccess propose également un export en MARC21.

Hyku vise à des applications plus étendues : il propose donc, en plus des formats classiques, un export vers EndNote (outil de gestion bibliographique non ouvert). Pour les futurs développements, des exports vers Zotero et Mendeley sont prévus.

- Alignement des autorités

Du fait du tropisme américain, les référentiels avec lesquels les logiciels s'alignent sont souvent ceux de la Bibliothèque du Congrès.

Hyku propose également un alignement avec GeoNames pour les toponymes.

Hyku, comme ArchivesSpace, travaille actuellement à développer d'autres alignements avec des référentiels existants.

4. USAGES ET ÉVOLUTIONS

4.1 Environnement linguistique

Développés en Amérique du Nord, les logiciels étudiés ont d'abord été réalisés en anglais. Ils s'orientent progressivement vers la traduction des interfaces dans d'autres langues. AtoM est disponible en français, Hyku en espagnol et en chinois, ArchivesSpace en français, allemand, espagnol et japonais, CollectiveAccess travaille actuellement à traduire son logiciel dans plusieurs langues.

Les logiciels étudiés semblent tous en mesure de produire des notices multilingues pour la description des fonds.

ArchivesSpace signale toutefois quelques difficultés dans le traitement en export des notices multilingues.

5. SYNTHÈSE

	EAD	Export	Briques logicielles	Hébergement	Communauté	Financement
AtoM	Oui, mais lacunes identifiées	CSV, XML	LAMP	Oui	Plusieurs centaines d'institutions internationales	Artefactual + Fondation en création
ArchivesSpace	Oui	CSV, XML, MARC	Apache, Ruby on Rails, MySQL	Oui	>450 membres (surtout Amérique du Nord)	Cotisation des membres institutionnels
CollectiveAccess	Oui	CSV, XML	LAMP	Oui	Plusieurs centaines d'institutions internationales	Conseil et développement par Whirl-i-Gig
Hyku	Oui	CSV, XML, EndNote, (Zotero)	Apache, Rails, Blacklight	Oui (via Fedora)	Très réduite	Partenariat institutionnel + fonds caritatif

BIBLIOGRAPHIE

Kelcy Shepherd, Bradley D. Westbrook, Lee Mandell et al., *The Archivist's Toolkit: An Integrated System for Describing and Managing Archival Resources*, 2006 (https://www.jstage.jst.go.jp/article/jcul/77/0/77_1216/_pdf)

Hanna Rosen, Jill Grogg, *LYRASIS 2021 Open Source Software Survey Report. Understanding the Landscape of Open Source Software Support for American Libraries*, 2021 (https://research.lyrasis.org/server/api/core/bitstreams/ac3f912c-cc6b-4745-8b70-2f36efb1f91c/content)

Digital Preservation Coalition, *Computational Access: A Beginner's Guide for Digital Preservation Practitioners*, 2022. https://www.dpconline.org/digipres/implement-digipres/computational-access-guide