arbido

2018/2 Automatisierung: Versprechen oder Drohung?

Cazeaux Hugues, responsable du pôle e-Research à l'Université de Genève

Krause-Bilvin Jan, Data specialist chez Docuteam Sàrl

Burgi Pierre-Yves, directeur du projet DLCM et directeur SI adjoint à l'Université de Genève

Tags: Archiv Elektronisch, Konservierung,

Automatisation de la préservation de données dans le contexte académique

Les universités n'ont pas attendu les directives du Fonds National Suisse (FNS) pour se pencher sur la question de la préservation et de l'accessibilité des données scientifiques. Hors du milieu académique, d'autres acteurs, comme Docuteam Sàrl, se sont aussi spécialisés dans cette problématique.

Préservation des données de la recherche

Lancé en septembre 2015 sous l'égide de Swissuniversities (association qui regroupe les responsables des hautes écoles universitaires, spécialisées et pédagogiques de Suisse depuis 2012), le projet *Data Life-Cycle Management* (DLCM)1 s'est fixé comme objectif d'élaborer une stratégie nationale en matière de préservation et de partage des données compatible avec les principes FAIR. Ces derniers permettent de s'assurer que les données soient trouvables, accessibles, interopérables et réutilisables.

Dans d'autres contextes (p.ex. des archives historiques), cette problématique est au goût du jour depuis des années. Docuteam Sàrl, spécialisée dans la gestion et la sauvegarde de l'information, s'est engagée dans ces problématiques depuis une dizaine d'années via la mise en place de standards2 et d'outils adéquats3, installés autant sur les infrastructures propres de ses clients que dans le cloud4. L'un de ces outils, docuteam feeder, est un moteur de workflow spécialisé dans l'archivage numérique. En d'autres termes, il s'agit d'un gestionnaire de tâches d'archivage.

Dans le domaine universitaire, les nouvelles réglementations évoquées ci-dessus ajoutent du travail supplémentaire aux chercheurs, dont la priorité est de publier leur recherche et non de gérer leurs données à des fins de préservation. Par conséquent, l'automatisation est un des axes sur lequel le projet DLCM a misé afin de faciliter l'adoption des bonnes pratiques de l'archivage tout en minimisant l'effort des chercheurs dans cette activité. Un élément clé pour ce faire repose sur une solution logicielle avec son jeu d'APIs (*Application Programming Interfaces*), qui facilitent l'intégration avec les instruments de laboratoire.

Nous vous proposons ici de considérer une combinaison de ces APIs du projet DLCM avec l'outil docuteam feeder afin d'augmenter les possibilités d'automatisation de la gestion des données de la recherche.

Projet DLCM et docuteam feeder

La solution DLCM consiste en une architecture ouverte et modulaire destinée à la préservation long terme des données de la recherche, conforme à la norme OAIS (ISO 14721) et compatible avec un déploiement dans le cloud. Concrètement, les différents modules de la solution offrent une gamme de services permettant aux chercheurs de préparer leurs données en vue d'être préservées, à savoir: de les soumettre avec une étape de *pre-ingest* suivi d' *ingest*, de les stocker physiquement (*archival storage*), d'indexer les métadonnées (*data management*) et de pouvoir y accéder selon des droits spécifiques (*access*). Cet ensemble de services garantissent la mise en œuvre des bonnes pratiques du domaine: recherche de virus, détection de format, calcul de somme de contrôle, contrôle d'intégrité, réplication, etc. Ces services sont mis à disposition via des APIs, présentés selon des Web services de type REST (*Representational State Transfer*). En d'autres termes, ils sont normalisés et donc agnostiques à la technologie. L'outil docuteam feeder est mis en œuvre dans le contexte de différentes plateformes d'archivage numérique OAIS, telles que docuteam cosmos. Cependant, et contrairement aux moteurs de workflows génériques, docuteam feeder est axé sur la préservation numérique. Par exemple, les actions entreprises sur les données peuvent être soigneusement consignées selon le standard PREMIS[5], un incontournable dans ce domaine. Étant basé sur la norme OAIS, docuteam cosmos et la suite DLCM présentent nombre de similitudes, telles que la modularité, les services en mode REST, etc. Ainsi docuteam feeder s'intègre naturellement dans cette suite d'automatisation en permettant d'organiser des tâches (voir Figure 1), telles que vérifier l'intégrité des données reçues, convertir les métadonnées d'un format vers un autre, s'assurer de l'absence de virus, reconnaître et migrer des formats spécifiques vers leur pendants mieux adaptés à la préservation à long terme (p.ex. des images PNG en images TIFF, des documents Word en PDF-A1, etc.). L'introduction de tels workflows s'avère précieuse dans le contexte de la gestion des données de recherche, en particulier par l'usage des modules d' *ingest* et de *data management*. En effet, il est important de conditionner au mieux les données avant leur stockage. Mais, comme évoqué précédemment, les chercheurs n'ont souvent que peu de temps à consacrer au conditionnement de leur information et toute automatisation représente un bénéfice précieux. Concernant le module *data management* (workflows de préservation), il devient incontournable face à l'évolution sans fin des technologies qui conduit à l'obsolescence des formats numériques. Son exécution périodique et automatique permet de se prémunir de ce type d'obsolescence, qui nécessite une planification de migrations en masse des formats (convertir les images en format JPEG2000, par exemple).

En conclusion, la plateforme DLCM, par sa flexibilité et facilité d'intégration aux environnements de travail du chercheur, s'impose comme une structure novatrice en matière de préservation appliquée au domaine académique. Cependant, concernant d'autres aspects de préservation (i.e., workflows d'automatisation), la solution docuteam feeder permet d'éviter de réinventer la roue et confère à la plateforme une expertise du domaine archivistique très précieuse.

1 Burgi, P.-Y., Blumer, E., Makhlouf-Shabou, B.. Research data management in Switzerland: National efforts to guarantee the sustainability of research outputs. IFLA Journal, 2017, p. 1-17, DOI 10.1177/0340035216678238 and «Home:: DLCM» [en-ligne] 2018. Consulté le 4.6.2018.

2 «Metadata Encoding and Transmission Standard (METS)» [en-ligne] 2018. Consulté le 4.6.2018.

3 «docuteam wiki» [en-ligne] 2018. Consulté le 4.6.2018.

4 «docuteam cosmos: solution cloud» [en-ligne] 2018. Consulté le 4.6.2018.

5 «PREMIS: Preservation Metadata Maintenance Activity» [en-ligne] 2018. Consulté le 4.6.2018.

Hugues Cazeaux

Après une formation d'ingénieur, Hugues a travaillé chez différents éditeurs logiciels pendant les 20 dernières années, en mettant en place des méthodes agiles de développement. Durant la dernière décennie, Hugues a acquis une grande expertise dans le domaine de l'archivage électronique et du Record Management. Actuellement, il s'occupe de définir et de développer la solution de préservation à long terme dans le projet DLCM. Il est aussi en charge du pôle *e-Research* à l'Université de Genève dont le rôle est de supporter les chercheurs de l'institution, avec comme mission principale de mettre en œuvre une plateforme de préservation des données de la recherche.

Jan Krause-Bilvin

Scientifique de formation, Jan a travaillé pendant une douzaine d'années au sein de diverses grandes institutions académiques et internationales dans le domaine de la gestion de l'information. Il est aujourd'hui employé chez docuteam dans divers projets de migration de données vers des systèmes de préservation numérique (docuteam cosmos et AtoM en particulier). Jan est également impliqué dans des mandats de consulting touchant à la mise en place d'infrastructures OAIS et enseigne la programmation à la Haute École de Gestion de Genève.

Pierre-Yves Burgi

Pierre-Yves Burgi à la responsabilité conjointe du service Solution Intégration et Développement (SoLID) de la division du Système de l'Information de l'Université de Genève. Il a reçu son diplôme d'ingénieur en informatique de l'École Polytechnique Fédérale de Lausanne en 1986, et le titre de docteur es science de l'Université de Genève en 1992, suivi d'un post-doctorat en neurosciences à San Francisco, USA. De 1997 à 2003, il a travaillé dans le département de la division micro-électronique du Centre Suisse d'Électronique et Microtechnique, Neuchâtel, où il a conduit des travaux de recherche appliquée dans le domaine de la vision artificielle sur la base de micro-circuits VLSI. De 2003 à 2017, il a dirigé le service des nouvelles technologies (NTICE) de l'Université de Genève. Ses intérêts incluent l'innovation dans l'enseignement, les humanités numériques et la gestion des données de

recherche.

Abstract

Français

Le projet national DLCM répond aux besoins des chercheurs en matière de préservation numérique, généralement exigée par les bailleurs de fonds (e.g. FNS). Mais les chercheurs se focalisent principalement sur leur recherche avec très peu de temps consacré à l'archivage pérenne de leurs données. D'où un besoin accru d'automatisation. Dans cet article, nous considérons la mise en commun de solutions respectant la norme OAIS, proposées par le projet DLCM et Docuteam Sàrl pour atteindre cet objectif.