Numéros / n° 2 - automne 2012

« Préservation du patrimoine numérique : manifeste pour la création d'un réseau de compétences »

Jérôme Barthélemy et Daniel Teruggi

Résumé

La préservation des données numériques est une problématique en soi, qui a donné lieu à des travaux spécifiques depuis plusieurs décennies. Nous en présentons un bref historique, et nous essayons de brosser un tableau du paysage actuel. Nous présentons notamment quelques problématiques non résolues, quelques approches particulières, bien que notre tableau soit forcément incomplet eu égard à la complexité du domaine. Nous terminons par un constat, et un appel.

Introduction

Après plusieurs décennies d'expérience, il apparaît évident que la numérisation massive de notre environnement, et notamment des données patrimoniales, a permis des progrès majeurs. En matière de diffusion de données, de mise à disposition pour les utilisateurs, la numérisation a permis à l'évidence un saut qualitatif et quantitatif, en mettant à disposition du plus grand nombre ce qui n'était à l'origine disponible que de manière parcimonieuse, et en permettant l'émergence de nouveaux usages. Dans le secteur culturel notamment, la mise à disposition des fonds, initiée par les moteurs de recherche, se poursuit grâce à des initiatives telles que la bibliothèque numérique européenne Europeana, la bibliothèque Gallica à la BNF ou le site Inamediapro de l'INA. De nouveaux usages de partage, enrichissements, « folksonomies » et autres pratiques collaboratives ont émergé et ont transformé notre approche des objets culturels, et en retour sont susceptibles de modifier des processus bien établis et normalisés tels que la veille informationnelle (Pirolli, 2011).

Toutefois, dans certains domaines, la numérisation massive a eu des conséquences dont peut-être nous ne mesurons pas toutes les retombées. Un domaine bien connu aujourd'hui est celui de la gestion des droits, mais un autre domaine de ce type est celui de la pérennisation à long terme du patrimoine, et de la création numérique sous toutes ses formes.

Bien que, à ses débuts, l'archivage numérique ait pu paraître comme une solution au problème de l'archivage à long terme, en raison de la possibilité de recopie à l'identique d'un contenu numérique, l'archivage numérique est rapidement apparu comme présentant des problématiques spécifiques dont la solution est loin d'être évidente.

1. Historique : la norme OAIS

Les agences spatiales, telles que le CCSDS [(1)], l'ESA [(2)], ou le CNES [(3)], ont été parmi les premières organisations à traiter du sujet de la pérennisation de l'information numérique. En effet, les données issues des satellites et des sondes spatiales sont numériques dès l'origine, et depuis près de 50 ans ? ce qui signifie notamment qu'elles ne sont pas le résultat d'un processus de numérisation, et qu'il n'existe donc pas de document original sous une autre forme. Tout document issu des observations des satellites est le

résultat d'un processus d'analyse des données, et d'un processus de rendu, qui sont eux-mêmes numériques. Le document résultant d'un processus d'analyse et de rendu ne peut donc remplacer le document original, et il devient dès lors indispensable de conserver les données originelles sous forme numérique, ce qui tranche par rapport aux techniques d'archivage sur bande magnétique ou microfilm.

1. 1. La norme OAIS

Les travaux effectués par ces agences les ont amenés à développer une norme, la norme OAIS (Open Archive Information System, ISO 14721 :2003), qui fait référence et qui s'impose aujourd'hui à tous les acteurs qui oeuvrent à la pérennnisation de données numériques.

Il importe bien de préciser que la norme OAIS n'adresse pas la problématique du support physique, mais celle du modèle de données et du modèle organisationnel qui doit être mis en place pour assurer la pérennité à long terme.

Cette norme a été mise en oeuvre dans de nombreux projets, tels par exemple que le projet européen CASPAR (4), qui a appliqué cette norme sur trois bancs de test différents : données scientifiques, données culturelles, et données de la création numérique contemporaine. Ce projet a donc montré que le modèle était parfaitement applicable dans de nombreux domaines, ouvrant la voie à des solutions techniques industrialisées ? plusieurs offres techniques sont aujourd'hui disponibles en provenance du secteur privé.

1. 2. La certification des services d'archive

Des travaux ont été entrepris plus récemment, visant à développer une méthodologie permettant de certifier la conformité d'un service d'archives avec les recommandations émises par la communauté de la recherche en préservation de données numériques. La méthode TRAC (5) (Trustworthy Repositories Audit and Certification : Criteria and Checklist), publiée en 2007, décrit près d'une centaine de critères à évaluer pour permettre la certification d'un service d'archives. Le catalogue Nestor (6) décrit de manière similaire une liste de critères d'évaluation, tout en les mettant en perspective pour une communauté spécifique, de langue allemande, et en l'appuyant sur un ensemble de bonnes pratiques spécifiques à cette communauté. La boîte à outils DRAMBORA (7) (Digital Repository Audit Method Based on Risk Assessment), a été développée pour permettre un contrôle interne des services d'archives, et est développée actuellement par le projet SHAMAN (8) (Innocenti *et al.*, 2009).

2. Problématiques non résolues

Malgré tous ces travaux, et bien que des sociétés commerciales se soient maintenant lancées dans la diffusion de services ou de solutions logicielles au problème de la préservation de données numériques, il reste toutefois que certaines problématiques d'archivage ne sont pas totalement résolues. Nous abordons ci-dessous, à titre d'exemple, la problématique de stockage, celle de l'identification pérenne des données, ou celle de l'authenticité.

2. 1. Stockage des données

En matière de stockage de données, la plupart des institutions mémorielles, telles que la BNF ou l'INA, ont mis en place une stratégie basée sur le stockage en ligne sur disque magnétique, complétée par un stockage sur bande, et procèdent à des migrations régulières du support physique. Toutefois, une telle stratégie est difficilement applicable par des particuliers, ou par des organismes qui génèrent de faibles volumes de données. Pour de tels usages, un stockage sur disque optique numérique pérenne serait une solution, et c'est d'ailleurs celle-ci qui est souvent utilisée. Toutefois, la pérennité des disques optiques numériques est faible, ce dont d'ailleurs la plupart des particuliers ne sont pas conscients, et le risque de perte de données patrimoniales est très important. Le développement d'un disque optique numérique pérenne reste donc une priorité, ainsi que le préconise le rapport du groupe PSN (pérennité des supports

numériques) commun aux deux Académies, des sciences et des technologies (Hourcade, Spitz et Laloë, 2010).

2. 2. Identification pérenne

La problématique d'identification pérenne demeure une problématique non résolue : comment s'assurer aujourd'hui qu'un identifiant affecté à un objet d'information, qu'il soit ou non numérique, puisse demeurer pérenne dans le temps ? Elle est largement reconnue, et fait partie des recommandations publiées sous le nom de « The New Renaissance », élaborées par le « comité des sages », groupe de réflexion sur la numérisation du patrimoine culturel européen, recommandations publiées le 10 janvier 2011 (9). On constate que dans certains domaines, comme par exemple l'identification des personnes, des standards sont en cours d'élaboration, comme le standard ISNI (10).

2. 3. Authenticité

Une autre problématique concerne celle de l'authenticité. La première version de la norme OAIS base l'authenticité sur deux concepts : l'identité de l'objet numérique d'une part, et l'intégrité du train de bits de l'autre. L'intégrité du train de bits est généralement assurée par une méthode de type « checksum », ou somme de contrôle, qui consiste à ajouter aux données d'autres données dépendant des premières, via une méthode de calcul connue et simple à appliquer. Toutefois, étant donné que la migration des formats de données est un passage obligé pour la plupart des données numériques que nous produisons, il est extrêmement difficile, et même impossible, d'assurer l'authenticité des données au sens premier de la norme OAIS, et notamment l'intégrité des données. Les concepteurs du modèle OAIS ont reconnu ce problème (Giaretta, 2009), et ont introduit la notion de « propriétés signifiantes » : l'authenticité serait assurée par la permanence de telles propriétés. Il reste évidemment une question essentielle : pour définir les propriétés signifiantes dans un domaine, il faut s'assurer que ces propriétés font consensus.

3. La collecte et la préparation des données

Toutefois, une problématique majeure demeure, qui ne peut être résolue par des services externes ou des solutions techniques, qui est celle de la collecte et de la préparation des données. Cette phase préalable à l'archivage ne peut être résolue de manière automatique, et demeure une problématique majeure de l'archivage numérique (Huc, 2004). Dans ce domaine, plusieurs approches différentes peuvent être tentées, nous détaillons ci-dessous certaines d'entre elles.

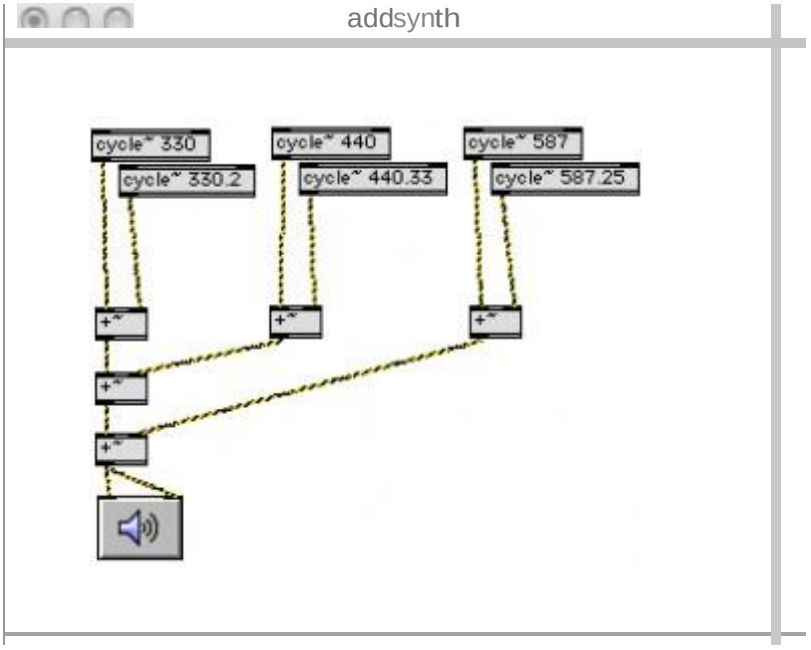
3. 1. Le projet ASTREE : production d'une documentation en langage naturel et notation mathématique

Le projet ASTREE (11) (partenaires IRCAM, MINES ParisTech, GRAME, université de Saint-Étienne) s'est proposé de développer une méthode spécifique pour la préservation à long terme des processus temps réel utilisés dans la création contemporaine, et notamment dans la création musicale (Bonardi, 2011 ; Barthélemy *et al.*, 2009). Il n'entre pas dans les objectifs de cet article de décrire en détail les procédés et méthodes développés par le projet. Toutefois, l'un des résultats majeurs du projet réside dans la possibilité de génération d'une documentation décrivant le processus temps réel, ou du moins sa partie « synchrone », en utilisant uniquement la notation mathématique et le langage naturel. Pour cela, le projet se base sur le langage « FAUST », développé au Grame depuis près de 10 ans (Orlarey, Fober et Letz, 2009). Ce langage possède des caractéristiques propres, dont la description dépasse le cadre de cet article, et qui en font une plate-forme pour des outils d'auto-documentation.

Dans le cadre de ce projet, deux outils ont été développés : un outil de transcription de Max/MSP vers le langage FAUST d'une part, et un outil de génération de documentation mathématique (Barkati et Orlarey,

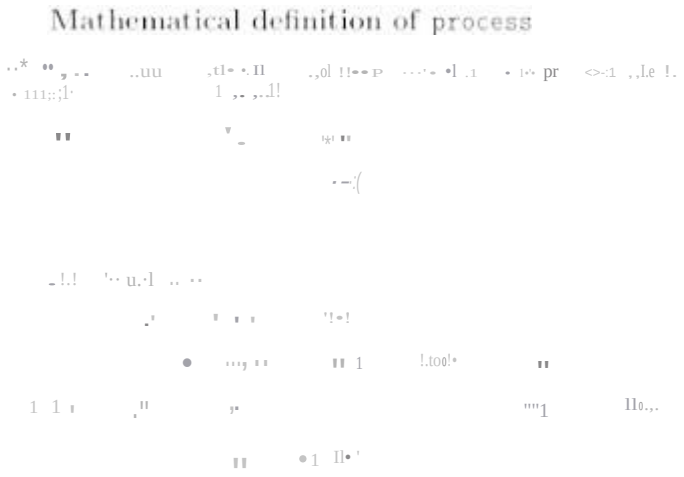
2011). Le résultat de ces opérations combinées permet de générer une documentation en langage naturel et notation mathématique des processus temps réel, comme illustré ci-dessous au travers dun exemple précis basé sur un processus de synthèse.

Figure 1



Un patch Max/MSP de synthèse additive

Figure 2



Un extrait (débu9 de la documentation générée autom a1iquem ent

La documentation rn athématique a été testée dans le cadre du Foi et ASTREE, en soumettant de telles documentations à deux ingénieurs infonn aticiens, qui ont réimplémenté les processus décrits en utilisant uniquement cette documentation, et sur des enviromements de programmation différents de Max&ASP (en l'occurence, Matlab et Mathematica). Ce test a permis de pointer quelques problèmes mineurs dans cette documentation, mais aussi un problème majeur : les deUle ingénieurs ont souhaité pouvoir disposer d'un jeu de données de test, qui n'est pas inclus dans la documentation à l'heure actuelle, afin de pouvoir valider leur développement. Toutefois, ce test valide globalement la possibilité de regénération du processus à partir de sa documentation.

3. 2. Le projet GAMELAN : tracer le processus de production

Le projet GAMELAN [(12)] (partenaires IRCAM, INA, EMI, UTC) se propose d'aborder la problématique de la production des connaissances sur les objets archivés en s'appuyant sur le processus de production, et notamment sur les traces laissées par les activités des utilisateurs. Dans le domaine de la production audio et musique, les objets générés sont des objets extrêmement complexes, comprenant souvent plusieurs milliers de fichiers de nature différentes et dont le rôle vis-à-vis de l'ensemble est mal défini. Pour des usages tels que la reprise ou la reproduction des oeuvres, ou leur « repurposing », ou bien pour des usages d'étude à vocation musicologique ou plus simplement didactique, l'archive du « master » n'est pas suffisante. Lors du processus d'archivage, qui intervient toujours à l'issue du processus de production, il est extrêmement difficile de réunir les informations nécessaires à la documentation des contenus. Cette situation amène de graves lacunes au niveau de l'information nécessaire, comme par exemple l'information complète sur les ayants droit. D'autre part, la documentation durant le processus de production est extrêmement difficile à réaliser, en raison du rôle spécifique des personnes en charge de la production.

Le projet GAMELAN se propose donc d'élaborer un méta-environnement de production, dont le rôle est de tracer les activités des utilisateurs, afin de permettre de replacer les éléments dans leur contexte, d'établir les liens entre les éléments et les actions effectuées par les utilisateurs. L'ensemble de cet environnement sera basé sur un modèle du processus de production audio et musique, développé dans le cadre du projet, et dont le développement relève du domaine de l'Ingénierie des connaissances (Vincent, Bonardi et Bachimont, 2011).

3. 3. Autres approches

Il existe bien d'autres projets de recherche, et notamment au niveau européen, visant à développer de nouvelles approches à la problématique de la préservation numérique. Deux de ces programmes, Digicurv et Arcomem, nous semblent mériter une attention particulière.

Le projet DigCurV [(13)], soutenu par la communauté européenne, se propose de définir un cursus de formation pour former des spécialistes de la préservation de données numériques, aptes à mieux comprendre les enjeux du numérique pour la préservation de données, et aider les producteur de données aussi bien que les services d'archives à s'adapter aux exigences du numérique. Le projet Arcomem [(14)], lui aussi soutenu par la communauté européenne, se propose de s'appuyer sur les réseaux sociaux ? et la « sagesse des foules » ? pour aider les services d'archives à sélectionner, enrichir, documenter et transmettre les connaissances au sujet des données numériques à préserver.

Conclusion : un constat, et un appel

Dans cet article, nous avons essayé de décrire le paysage de la préservation de données numériques. Notre description était forcément extrêmement partielle, mais nous avons pu voir que ce paysage était extrêmement divers : les compétences en la matière vont des compétences en matériaux et procédés (pour les supports physiques), à l'ingénierie des connaissances, en passant par les langages de programmation, les techniques de collaboration en réseaux, les sciences de l'éducation, pour ne pas parler des disciplines traditionnelles de l'archivistique, telle que la diplomatique, dont l'apport ne pourra pas être négligé. D'autre part, des solutions commerciales commencent à émerger, mais leur mise en oeuvre nécessite des compétences spécifiques : ce ne sont en aucun cas des « boîtes noires » qu'il suffirait d'invoquer pour permettre la préservation à long terme des données qu'on leur confierait.

Ceci nous amène à un constat : la diversité des compétences à mettre en oeuvre interdit à l'évidence à un acteur isolé toute possibilité d'élaboration d'une solution complète à la préservation des données numériques qu'il produit, sauf bien entendu lorsque cette mission est au coeur de son activité, comme c'est le cas en France pour les grandes institutions mémorielles telles que l'INA ou la BNF. Pour les autres

acteurs, producteurs de données numériques dans de nombreux domaines, ils ne peuvent pour l'instant s'appuyer sur aucune structure susceptible de leur fournir les services ou les compétences adéquates.

Dans certains domaines, ce constat a déjà été effectué, et des réseaux de compétences et des services ont été mis en oeuvre pour aider à la solution de ce problème.

Dans le secteur des sciences humaines et sociales, la France a mis en place le Très Grand Équipement Adonis (15), qui permet aux producteurs de données numériques d'avoir accès à une grille de services permettant l'archivage et garantissant un accès pérenne aux documents issus de la recherche en SHS.

Les acteurs du secteur scientifique ont mis en place, au niveau européen, le projet APARSEN (16). Celui-ci pose les bases d'une infrastructure garantissant un accès pérenne à l'information et aux données numériques.

Dans le secteur audiovisuel, plusieurs acteurs ont mis en place le centre de compétences PrestoCentre (17), dont l'objectif est de fournir des compétences en matière de préservation à long terme aux acteurs du secteur.

Il nous semble donc qu'une étape cruciale pour les producteurs de données numériques dans la création artistique contemporaine devrait passer par la mise en place d'une telle structure. Celle-ci devrait proposer des services et des compétences, collecter l'information sur les bonnes pratiques et les nouvelles approches, et diffuser ces connaissances dans la communauté de la création artistique contemporaine. Elle pourra s'appuyer sur les compétences et les outils mis en oeuvre dans d'autres domaines, comme par exemple les registres de formats de données, les standards, notamment en matière d'identification, les procédures de certification, les outils commerciaux existants. Elle pourra initier les travaux de recherche nécessaires dans les domaines actuellement non couverts, par exemple en matière de notation de l'interaction. Il n'entre pas dans les objectifs de cet article de décrire en détail l'organisation d'une telle structure, qui reste à créer. Notre propos ne vise qu'à appeler au rassemblement des voix et des volontés pour faire aboutir un tel projet.

---

1. Consultative Committee for Space Data Systems
2. European Space Agency
3. Centre national d'études spatiales
4. http://www.casparpreserves.eu
5. http://www.dcc.ac.uk/resources/tools-and-applications/trustworthy-repositories
6. http://files.d-nb.de/nestor/materialien/nestor_mat_08-eng.pdf
7. http://www.repositoryaudit.eu/
8. http://shaman-ip.eu/shaman/
9. http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/final_report_cds.pdf
10. ISNI: "draft ISO standard" ISO 27729, www.isni.org
11. Le projet ASTREE est un projet soutenu et financé par l'Agence nationale de la recherche.

12. Le projet Gamelan est un projet soutenu et financé par l'Agence nationale de la recherche.

13. http://www.digcur-education.org/

14. http://www.arcomem.eu/

15. http://www.tge-adonis.fr

16. http://www.alliancepermanentaccess.org

17. http://www.prestocentre.org

Pour citer ce document:
Jérôme Barthélemy, « Préservation du patrimoine numérique : manifeste pour la création d'un réseau de compétences », *RFIM* [En ligne], Numéros, n° 2 - automne 2012, Mis à jour le 28/09/2012
URL: http://revues.mshparisnord.org/rfim/index.php?id=193