Retour d'expérience de ma participation à la conférence iPRES22

Pierre-Yves Burgi
pierre-yves.burgi@unige.ch
https://orcid.org/0000-0002-4956-9279
Deputy CIO, Université de Genève

Résumé

Organisée par la Digital Preservation Coalition (DPC), la 18ème conférence internationale sur la préservation numérique (iPRES) a eu lieu à Glasgow, en Écosse, du 12 au 16 septembre 2022. Cette édition correspond également au 20ème anniversaire du DPC, une organisation ancrée à Glasgow, qui a voulu de ce fait accueillir des participants du monde entier dans leur ville. Le résultat a été une conférence hybride avec 649 participants qui se sont joints en personne (437) et en ligne (212).

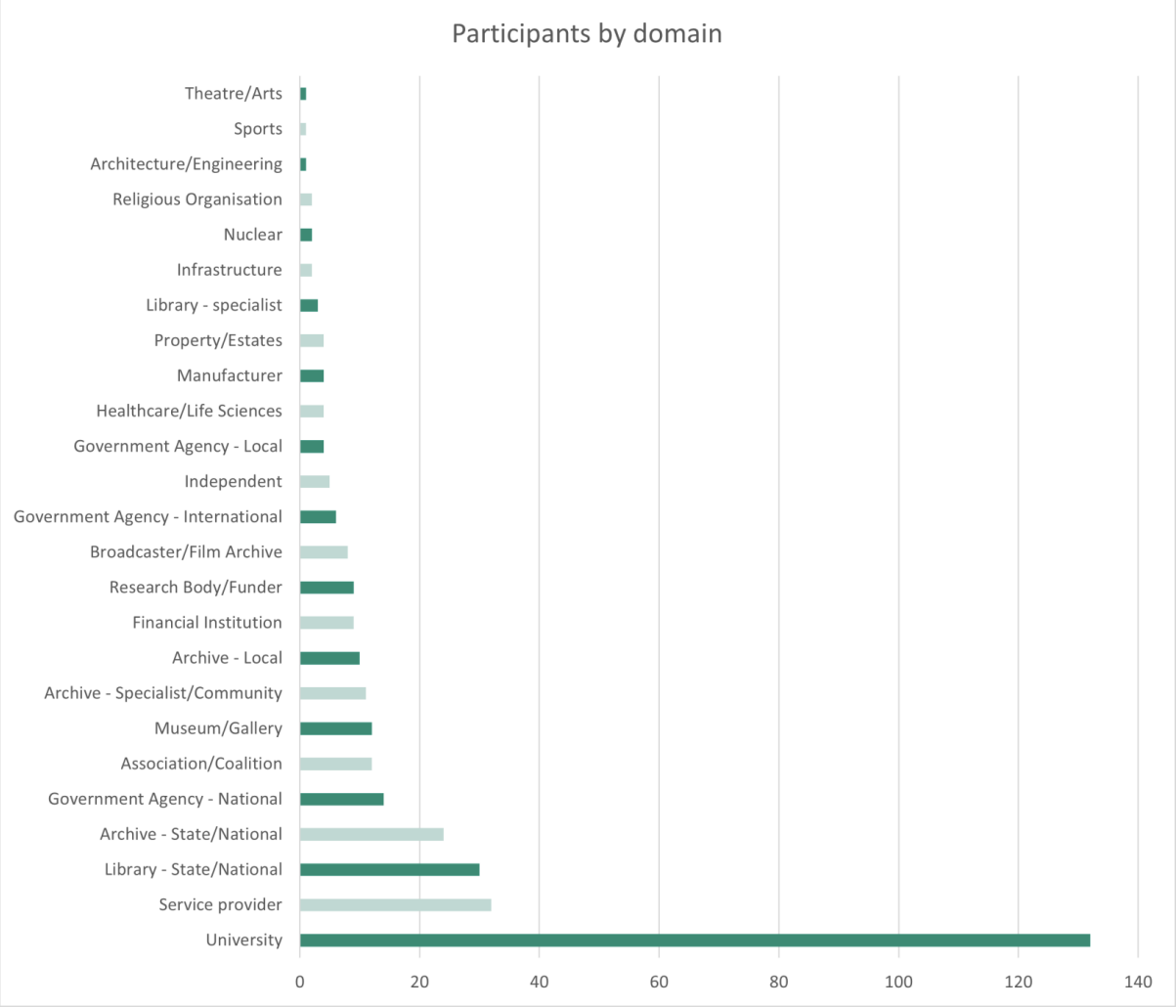
Mots-clés

iPRES22, préservation numérique, OLOS, DLCM, Stockage ADN, e-ARK

1. Introduction

Organisée par la Digital Preservation Coalition (DPC), la 18ème conférence internationale sur la préservation numérique (iPRES) a eu lieu à Glasgow, en Écosse, du 12 au 16 septembre 2022. Cette édition correspond également au 20ème anniversaire du DPC, une organisation ancrée à Glasgow, qui a voulu de ce fait accueillir des participants du monde entier dans leur ville. Le résultat a été une conférence hybride avec 649 participants qui se sont joints en personne (437) et en ligne (212). La Figure 1 illustre le large éventail de secteurs professionnels liés à la préservation numérique, ce qui témoigne de la diversification de la communauté.

Figure 1 : Éventails des secteurs professionnels représentés à iPRES22 (source : the iPRES 2022 Proceedings)



iPRES 2022 s'est tenue sous la thématique "Des données pour tous, pour le bien et pour toujours". L'appel à contributions invitait à la réflexion et au débat sur la manière dont la préservation numérique peut soutenir des communautés, des écologies, des économies et des idées florissantes, et il s'articulait autour de moments et d'idées de l'histoire de la ville. Les organisateurs ont également voulu faire adopter la devise de la ville hôte, "Let Glasgow Flourish", avec le sous-titre "Let digits flourish". Il s'agit d'un jeu de mots très « scottish » sachant que le nom "Glasgow" signifie littéralement "Cher endroit vert" !

Selon la tradition iPRES, le lundi est dédié à des tutoriaux et ateliers, les 3 jours suivants à la conférence principale, puis le vendredi les participants ont été invités à assister à l'une des 14 visites professionnelles organisées dans des institutions à travers l'Écosse. C'est dans ce contexte que j'ai participé à cette conférence et eu l'occasion de présenter avec mes collègues d'une part la technologie DLCM, et d'autre part l'usage potentiel de l'ADN pour archiver les données au travers d'une présentation et d'une table ronde sur cette thématique.

Cette année les présentations ont été réparties selon 6 catégories : environnement, innovation, résilience, communauté, jeux et échange. Pour ma part, je me suis essentiellement concentré sur les quatre premiers, voir ci-dessous. Vu le nombre de sessions il fallait bien faire un choix, bien que le domaine de l'émulation d'anciens jeux soit tout aussi passionnant. En effet, ce domaine met en évidence des OS disparus et des formats de fichiers obsolètes, une problématique qui nécessite des solutions technologiques très pointues, innovatives et efficientes qui ont des retombées dans d'autres domaines (donc celui de l'«échange»).

2. Environnement

L'impact environnemental de la préservation numérique, consommation électrique, bilan CO2, durabilité, etc., a été abordé à plusieurs occasions :

- The C02 emissions of Storage and use of Digital Objects and Data: Exploring Climate Actions and Impact measures to consider.
- Green goes with anything.
- Seeking sustainability: Developing a Modern Distributed Digital Preservation system
- The climate crisis and new paradigms for digital access.

C'est clairement un sujet complexe qui néanmoins commence à être traité avec une granularité très fine. Par exemple l'impact sur la consommation énergétique a été évalué sur des cas concrets :

- du nombre de fois dans une année le calcul des checksums est effectué ;
- du nombre de copies des archives dans les data centers ;
- l'emplacement de ces copies sur des serveurs locaux, dans le cloud, distribué selon l'architecture LOCKSS, etc.

Au-delà de l'énergie, une présentation traitant des aspects liés aux chaînes d'approvisionnement agroécologiques démontre le potentiel du cycle de vie des données. Tyfu Dyfi - Food, Nature and Wellbeing (alimentation, nature et bien-être) est un projet pilote qui soutient et développe l'agroécologie dans une réserve de biosphère soutenue par l'UNESCO. Financé par le programme ENRaW (Enabling Natural Resources and Well-being) du gouvernement Welsh, il rassemble une série de partenaires et de producteurs utilisant des méthodes agroécologiques pour démontrer "comment les communautés peuvent être impliquées dans leurs systèmes alimentaires locaux et énumérer les multiples avantages qui en découlent". Un espace d'information basé sur une modélisation robuste des données, et une architecture de données qui englobe le cycle de vie complet, de la création à la conservation post-création, l'accès et l'utilisation, optimise la mise en relation des producteurs et consommateurs.

3. Innovation

3.1. Stockage ADN

Cette année, et c'est une première, une session entière a été dédiée au stockage de l'information dans l'ADN, avec d'une part ma contribution et celle de mes collègues : « OAIS-compliant digital archiving of research and patrimonial data in DNA », suivi d'une autre présentation complémentaire « DNA Data storage for long term digital preservation ». Cette session s'est clôturée avec une table ronde « Will DNA form the fabric of our digital preservation storage? » à laquelle j'ai participé aux côtés de représentants de l'industrie du domaine (Twist Bioscience), du DPC, et de l'académique (universités de Yale et de Californie). L'auditoire fut rempli avec ~140 personnes, avec un public très engagé qui a posé d'excellentes questions aux panélistes, ce qui démontre un intérêt certain pour cette technologie émergente. Parmi les questions, il est ressorti que les archivistes n'ont pas une grande confiance dans les supports Write-Once en raison de technologies passées qui n'ont pas su démontrer leur résilience. Ils ont d'autre part des exigences spécifiques, notamment en matière de métadonnées et de migrations des données.

3.2. Projets européens : ARCHIVER et e-ARK

Le projet ARCHIVER (Archiving and Preservation for Research Environments) a passé plus de trois ans à concevoir, prototyper et piloter de nouveaux services innovants pour la préservation numérique à long terme de données scientifiques. Au cours du projet, plusieurs organisations producteurs de grands volumes de données représentant plusieurs domaines de recherche (CERN, DESY, PIC et EMBL-EBI) ont travaillé en étroite collaboration avec des fournisseurs de solution (Arkivum et LIBNOVA) sur la recherche et le développement de nouveaux services et solutions pour la préservation des données scientifiques pertinentes pour l'European Open Science Cloud (EOSC). Lors d'une table ronde, les acteurs du projet ARCHIVER ont partagé leur expérience et les leçons apprises au cours du projet. Les sujets abordés ont été les suivants :

- les avantages d'une approche collaborative entre les utilisateurs finaux et les fournisseurs commerciaux ;
- les défis qui ont été relevés en cours de route et les solutions qui ont été créées ;
- ce qu'il reste à faire pour concrétiser la vision du projet, à savoir des services de conservation numérique durables pour l'ensemble de la communauté scientifique, qui répondent aux besoins des organisations de toutes tailles.

Cette session intervient 2 mois seulement après la fin de la phase pilote finale du projet ARCHIVER, ce qui en fait le moment idéal pour partager leurs idées et leurs expériences.

Le Consortium E-ARK a travaillé ses 10 dernières années sur des spécifications, des outils et des bonnes pratiques pour l'archivage numérique en Europe et au-delà. Le portefeuille d'exemples de logiciels E-ARK comprend deux systèmes d'archivage numérique matures, RODA de KEEP Solutions et ESS Arch de ES Solutions. ES Solutions est une société suédoise spécialisée dans la gouvernance de l'information. ESS Arch, tout comme RODA, est basé sur le modèle OAIS (Open Archival Information System, ISO 14721:2003) et complété pour y inclure les fonctions de PreIngest et PreAccess. Cette présentation fut l'occasion de

montrer des exemples concrets d'institutions qui ont adopté e-ARK. A noter que d'une part les Archives fédérales suisses ont participé à E-ARK et leur collaboration a contribué de manière substantielle au développement du format SIARD et d'autre part des concepts de E-ARK ont été utilisés lors du projet swissuniversity DLCM, qui a conduit au développement de OLOS.

Une autre présentation d'intérêt, centrée sur RODA, a traité les différentes étapes du cycle de vie du contenu d'une archive, avec un focus particulier sur les règles de destruction (« disposal »), qui évidemment nécessitent des précautions particulières.

3.3. Internet Archive

Depuis sa création en 1996, l'Internet Archive, une bibliothèque à but non lucratif, fournit une infrastructure et des services de stockage, de préservation et d'accès à plus de 1000 organisations du patrimoine culturel dans le monde. Depuis sa création, Internet Archive s'est efforcé de garantir la disponibilité et l'accessibilité des connaissances humaines en créant une bibliothèque pour stocker de manière permanente le contenu numérique. Internet Archive a mis en place un nouveau service de préservation numérique à usage général pour compléter et étendre sa gamme existante de services gratuits, payants et à but non lucratif pour la numérisation, l'archivage Web, le stockage général des données et services Web. Le nouveau service de préservation numérique, appelé Vault, est construit sur l'infrastructure existante d'Internet Archive et sur des logiciels libres. Il a intégré les feedbacks de dizaines de partenaires pilotes et de pairs qui utilisent le service au fur et à mesure de son développement et de sa progression dans le cycle de vie du produit.

À la base, Vault permet aux utilisateurs de déposer n'importe quel contenu numérique, quelle que soit sa taille, de spécifier l'emplacement géographique où leurs données seront stockées (sur plusieurs sites, actuellement trois pays différents), de définir le nombre de copies des données qui seront répliquées et leur répartition entre différents centres de données dans diverses régions, de choisir s'ils veulent que leurs collections soient stockées selon différentes architectures et de sélectionner la fréquence des opérations d'audit et de vérification de la fixité des objets numériques.

3.4. Résilience

Plusieurs présentations ont traité du sujet de résilience, par exemple « The design and implementation of a necessary and sufficient system for the long-term archival retention of digital documents ». Cependant une présentation particulière a retenu mon attention, à savoir un sujet d'actualité : le décommissionnement des centrales nucléaires. Ce domaine repose sur l'archivage à très long-terme des données, sur des milliers d'années, par exemple pour assurer une gestion des fûts radioactifs. Pour cela une solution originale, « passive » sur papier a été présentée. Elle consiste à décrire l'algorithme qui permet de relire les données cryptées également sur papier, le tout tenant sur une centaine de pages (représentant un facteur de compression >103 en nombre de pages comparativement aux pratiques actuelles de documentation). Ce document a été distribué à des étudiants sans grandes compétences en informatique qui ont réussi après deux semaines à décoder toutes les données sur cette base (sans autres explications), une bonne démonstration du principe d'autodescription préconisé par la norme OAIS !

A noter que dans le même session, une autre présentation au titre intriguant "From ray cats to DPC RAM: How best to preserve a digital memory of the nuclear decommissioning process",

à savoir un nouveau modèle de maturité appelé RAM (Rapid Assessment Model), a été développé par le Nuclear Decommissioning Authority en collaboration avec le DPC.

4. Communauté

4.1. Formalisation des politiques sur la normalisation des formats de données

La BnF (Bibliothèque nationale de France) a présenté un sujet très controversé sur la manière de définir une politique sur les formats de données pour la conservation numérique. En effet, la BnF a dû formaliser une méthode pour traiter les données collectées qui ne répondaient pas à ses exigences. Plusieurs exemples concrets ont conduit la BnF à passer d'idées préconçues à des décisions pragmatiques sur les stratégies de normalisation et de préservation des contenus qui ne pouvaient pas être ingérés tels quels. L'intelligence collective a été fortement sollicitée entre experts, responsables de la collecte et responsables du processus. La démarche a été structurée selon trois questions : (1) Est-il nécessaire de transformer les données reçues ? Si oui, (2) dans quel format ? (3) Faut-il conserver les fichiers sources ?

Une fois ces décisions prises, le choix d'un format cible nécessite un examen plus approfondi, en utilisant trois critères pertinents pour ces cas d'utilisation et qui se conforme à leur politique:

1. catégorie de format : identification d'un format préféré pour le type de contenu concerné ;
2. cohérence au sein du paquet d'informations ou de la collection : identification des formats présents dans le paquet d'informations ou la collection, à privilégier en cas de formats préférés multiples ;
3. préservation des propriétés ou fonctionnalités significatives : définition d'une intention de préservation, c'est-à-dire l'ensemble des propriétés informationnelles et des modalités d'utilisation d'un objet numérique à préserver à long terme pour une communauté d'utilisateurs.

Le cas concret des formats d'images a particulièrement bien illustré la difficulté et controverse qui peut résulter de la démarche !

4.2. Archivage des emails

Dans la même session une présentation a traité de la problématique de l'archivage des emails. Le projet RATOM-FIRE (Review, Appraisal and Triage of Mail - Functional, Interoperability and Reuse Extensions) a l'intention de développer une suite d'outils pour traiter de manière fiable et efficace les collections de courriers électroniques. Le résultat du logiciel est conçu pour faciliter un large éventail d'activités de conservation. Le projet RATOM-FIRE répondra aux besoins identifiés par la communauté archivistique pour intégrer les résultats de l'outil dans les flux de travail de conservation numérique existants et émergents. Il s'agira notamment d'exporter plus facilement les messages électroniques sous forme de fichiers individuels (eml), de capturer des métadonnées de conservation plus détaillées et d'étendre l'interface de programmation d'application (API) pour faciliter l'intégration dans d'autres outils.

5. Keynote

Des trois *keynote speakers* qui sont intervenus durant les trois jours de conférence, Steven Gonzalez Monserrate, un ethnologue du « cloud », a particulièrement retenu mon attention de par l'originalité de son domaine d'étude. Sa thèse porte sur les divers impacts écologiques de l'informatique et du stockage des données numériques, en particulier des data centers en UK, en Arizona, à Porto Rico et en Islande. Il a abordé des problématiques spécifiques comme l'étude du comportement des ingénieurs qui s'occupent des serveurs en salle machine, ainsi que les conséquences nuisibles sur les populations qui vivent à proximité des data centers. Selon son étude, il ressort en effet que les exploitants de ces centres de calculs prennent à cœur de refroidir les serveurs au-delà de ce qui est nécessaire afin de se protéger contre les conséquences de dysfonctionnements qui dans certains cas pourraient mettre en péril leur poste. Un tel refroidissement impose des nuisances sonores dues à des compresseurs en perpétuels fonctionnement (avec aussi des conséquences sur le réchauffement climatique…). Parmi les sujets abordés, comme solutions aux problématiques actuelles, il a mentionné des projets de centres de données sous-marins ou extra-terrestres, les cristaux de mémoire 5d, les jardins de données alimentés par des capacités de stockage d'ADN synthétique et les nouvelles technologies de calcul quantique (cette dernière technologie utilisant aussi des moyens de refroidissement gourmands en énergie).

6. Conclusion

Depuis le COVID, iPRES 2022 fait partie des premières conférences en présentiel. Cela m'a rappelé combien le networking se fait lors des pauses et autour d'une table au restaurant le soir, et non pas au travers d'un écran ! Cela aura été la troisième fois que j'ai l'opportunité de participer à la conférence annuelle iPRES. Aussi, de ces participations, je ne peux que recommander à tous les professionnels du domaine cette conférence tant les présentations, tutoriels, ateliers, posters, etc. qui font partie du programme sont de grande qualité et traitent de sujets si divers que chacun y trouvera son intérêt. De plus, la communauté francophone (i.e., suisse, française, belge, canadienne, etc.) y est très bien représentée !

https://doi.org/10.55790/journals/ressi.2023.e1097