Aristote
À la croisée des révolutions numériques

Webinaire PIN
GoToWebinar

La certification d'entrepôts de données – panorama, évolutions et retours d'expérience

Jeudi 30 mars 2023

Coordination scientifique :

BnF

cnrs

Renseignements, programme…

https://www.association-aristote.fr/evenements/pin-la-certification-dentrepots-de-donnees/

Association Aristote : Groupe PIN

GoToWebinar

Compte-rendu de la réunion du 30 mars 2023

Thème :
La certification d'entrepôts de données – panorama, évolutions et retours d'expérience

33 personnes présentes.

- **Tour de table, présentation de l'ordre du jour, événements passés et futurs dans le domaine d'intérêt de PIN : formations, publications et rapports divers, colloques et ateliers.**

1. Publications, rapports, etc.

PIN/Cellule Formats : Traduction de deux notes d'orientation de veille technologique de la série Type de Données de la DPC :

- Préservation des systèmes d'information géographique (SIG)
- Préservation des feuilles de calcul.

Library of Congress (LOC) - Révision des formats recommandés pour l'archivage : https://www.loc.gov/preservation/resources/rfs/ - appel à commentaires ouvert jusqu'au 14/04

Digital Preservation Coalition (DPC) - Digital Preservation toolkits :

- Policy https://www.dpconline.org/digipres/implement-digipres/policy-toolkit
- Business case : https://www.dpconline.org/digipres/implement-digipres/business-case-toolkit

Open Preservation Foundation (OPF) - Publication Jhove 1.28 : https://openpreservation.org/news/first-jhove-1-28-release-candidate-out-now/?q=3

National Digital Stewardship Alliance (NDSA) - Enquête utilisation : https://docs.google.com/forms/d/e/1FAIpQLSfOlcI-QCYQ1V2504pIRxjAqpBvBxmDYlW7ZhuRis5HepQNrw/viewform - réponses jusqu'au 01/05

2. Conférences :

Durability of Digital Storage (DDS) : https://dds.hypotheses.org/ journée de lancement le 13/04 (Nancy, France)

Open Repositories 2023 : https://or2023.openrepositories.org/ (Stellenbosch, SA) 12-15/06

Archiving 2023 : https://www.imaging.org/IST/IST/Conferences/Archiving/Archiving2023/Archiving2023_Home.aspx (Oslo, Suède) 19-23/06

iPRES 2023 : https://ipres2023.us/ (Chicago, USA) 19-22/09
NDSA Digital Preservation : https://ndsa.org/conference/ (St Louis, USA) 15-16/11

3. Agenda Aristote / PIN :

Formation : 11-15/09 Prochaine plénière : 6/07
Séminaire : Data Centers, IA, Cloud, High Performance Computing : Quel impact environnemental ? Comment le mesurer et le contenir ?
https://www.association-aristote.fr/evenements/data-centers-ia-cloud-high-performance-computing/ le 13 avril.

- **Panorama général international CoreTrustSeal : historique, évolutions récentes et perspectives par Olivier Rouchon (CNRS)**

Revue des différentes initiatives de certification et d'auto-certification dont TRAC, World Data System…
Data Seal of Approval et Core Trust Seal ont des approches similaires même si elles touchent des communautés différentes.
En 2016, DSA + WDS devient CTS sous l'impulsion de la Research Data Alliance.
L'assemblée des évaluateurs constitue l'équipe qui assure la certification.

16 +1 critères d'évaluation organisées en 3 blocs :

- Organisation de l'infrastructure
- Gestion des objets numériques
- Système d'information & sécurité

Détails : https://doi.org/10.5281/zenodo.7051011
Basé sur OAIS : https://fr.wikipedia.org/wiki/Open_Archival_Information_System

Modèle de certification européen
simple : CoreTrustSeal
16 critères
170+ critères
étendue : nestorSeal DIN 31644 (norme allemande - langzeitarchivierung.de)
34 critères
4 centres
formelle : ISO 16363:2012
91 critères
2 centres
3 entrepôts français actuellement certifiés CTS
CDS (Strasbourg)
Ifremer
(Brest)IDOC
(Orsay)

Principes de bases du CTS :

1. Auto-évaluation sur la base d'un questionnaire (nécessite des compétences variées : managérial, archiviste, technique , …). Revient à définir le fonctionnement de l'entrepôt / archive OAIS de manière sourcée cad en amenant les preuves (documentations, diagrammes, textes réglementaires, site web…) de ces bonnespratiques ou ce bon fonctionnement.
2. Évaluation en aveugle par deux experts de l'assemblée des évaluateurs.

3. Évaluation des rapports par le board
4. Acceptation du dossier (nécessite jusqu'à 5 allers-retours entre le candidat et les évaluateurs)

Prévoir des frais administratifs et de gestion de 1000 €. Possibilité de se faire financer les frais pour la première candidature par le CoSO (Conseil pour la Science Ouverte).

Passage en revue des critères version 2023-2026.
Comparaison avec la norme 16363 d'audit d'un système OAIS qui a un découpage similaire en trois blocs mais en s'appuyant sur 96 critères (= plus complexe à mettre en oeuvre)

Qu'est-ce qu'un entrepôt de confiance ? = entrepôt fiable …

Avantages de cette certification : une première phase d'auto évaluation pour ne rien oublier (= pas d'angle mort, évaluation interne) + évaluation par ses pairs (évaluation externe).

Le plan national pour la science ouverte incite fortement à la certification - priorité nationale.

Montée en puissance de l'importance de la RGPD

Quantité de travail à prévoir = généralement quelques semaines pour une

équipe.Le CTS clarifie le paysage. Et il est international.

CTS est d'actualité notamment avec la montée de la science ouverte ; Pour autant, avec un équilibre économique fragile.

Il existe des ateliers de formation (RDA-France). Et le GT COSO/RDA-France peut prendre en charge les frais de première candidature.
https://listes.services.cnrs.fr/wws/subscribe/rda-france

A noter : Les réponses aux critères sont à faire en anglais ; les documents sources fournis en référence peuvent rester dans la langue du candidat.

Questions :

- les data centres peuvent-ils candidater ? C'est l'utilisation qui est certifiée - par exemple par la mise en œuvre d'une politique de préservation. C'est donc des organisations qui sont certifiées. Par exemple, Vitam du CINES ne sera pas certifié, mais ses applications.
- le CTS peut-il s'appliquer aux entrepôts privés ? Tout à fait.
- ANF 461 : il y a beaucoup plus de détails exigés ; et l'audit est très différents. LeCTS est basé sur la confiance, qui n'est pas mise en cause a priori.
- d'évaluations faut-il prévoir en 3 ans (pour 1 entrepôt certifié) ? Chaque entrepôt désigne un évaluateur. On n'évalue jamais seul, mais avec un senior. Et il existe des réunions annuelles. Il faut tabler sur 2 à 3 évaluations par an et par évaluateurs : 2 à 3 demi-journées de travail par évaluation.
- Existe-t-il des certifications en Chine ou au Japon ? Des entrepôts chinois et japonais sont certifiés. Mais l'Europe et l'Amérique du Nord constituent 70% des

certifiés.

- **Retours d'expérience CTS : SISMER/IFREMER Brest par Christine Coatanoan (IFREMER)**

SISMER gère les systèmes d'information scientifique de la mer. Il a un rôle central dans la gestion des bases de données marine à l'international. Il a été créé en 1989. Il doit répondre aux principes FAIR, et il est certifié par l'UNESCO.
ODATIS est un pôle de données et services pour l'océan. C'est l'un des 4 pôles de données sur la terre.

Pourquoi la certification ? pour la confiance en termes de fiabilité et de durabilité.

Deux cas différents : des données homogènes ; et des données hétérogènes.

Ensemble de services au service des producteurs et des utilisateurs de données marines :
- brique découverte des données
- plateforme d'analyse des données
- support (?)

IFREMER dans une démarche qualité ISO 9001 qui a permis de fournir des éléments réutilisés dans le cadre de la certification CTS.
A noter que les filières sont très hétérogènes.

Les sources de données sont diverses : allant de bouées dérivantes à des données satellitaires.

SISMER bénéficie de différentes certifications : par ex; IR Data Tera, iso 9001…

Participants à la certification variés : responsable du système IDM, responsable de processus (P8), …
La granularité répond à l'organisation interne en 6 filières.
Première soumission en mars 2019 pour une certification en novembre 2019.
https://www.coretrustseal.org/wp-content/uploads/2019/11/IFREMER-SISMER.pdf
En cours de renouvellement depuis le 20 octobre 2022.

Très intéressant de voir les améliorations apportées d'un renouvellement à l'autre, pour arriver au niveau 4 du critère. L'un des relecteurs était très précis dans son questionnement, ce qui a guidé les réponses. Être évaluateur montre d'autres manières de faire, et accroît l'expérience.

<u>Questions :</u>
- Pouvez-vous développer le critère sur les licences ? Pas mal de données étaient en accès restreints, sans être aux bonnes normes. Nous avons travaillé sur l'accessibilité extérieure. Ce n'est pas toujours facile de bien décrire chaque filière, et nous n'avons pas toujours été assez précis. Il faut bien expliquer comment sont gérées les données et leur accessibilité.
  Il est assez difficile d'avoir des choses précises pour tous les jeux de données - la

licence exacte qui s'applique, etc. Il faut rédiger un document sur la licence par défaut qui s'applique en absence de licence spécifique. Mais tous les relecteurs ne sont pas d'accord.

- Coût en ressources humaines ? N'ayant pas participé à la première rédaction, ce n'est pas facile à évaluer. Le renouvellement n'est pas trop compliqué, à condition debien travailler en équipe bien identifiée. Beaucoup de dossiers aident à la rédaction (rôle des guides). Mais s'inscrire dans un processus de certification exige de l'intégrer au quotidien pendant un certain temps pour que tout ce qui est mis en place soit fait dans l'optique de la certification. Cela a un coût supplémentaire pendant plusieurs années, qui est difficile à quantifier. A IFREMER, il existait des documents existants qui ont facilité la tâche. Donc maintenir un référentiel documentaire que l'oncollecte pour la certification.
- Coût de la mise à jour en termes de ressources humaines ? 4 ou 5 personnes onttravaillé sur le dossier, dont une qui a vérifié tous les liens, et une autre pour la relecture. De lordre de quelques semaines …

♦ Retours d'expérience CTS : CDS Strasbourg par Françoise Genova (CDS)

Centre des données astronomiques de Strasbourg. Il vient de fêter ses 50 ans. Il a une dimension internationale.
Sa mission est restée la même et reste d'actualité.
Continue à assurer la qualité et la curation des données pour servir la recherche.
Est une infrastructure de recherche confirmée depuis 2008.

La certification concerne deux services seulement : VizieR et Aladin.
Ce qui est certifié dans la galaxie des outils et systèmes du CDS est le système ALADIN.
ALADIN est un outil de visualisation du ciel & est également une collection d'images du ciel.

- Documentation des process de VizieR/Aladin : https://cds.u-strasbg.fr/vizier-org/CTS-2022/html/VizierProcesses_v1.2_16mar2022.html
- formalisme de l'OAIS pour les workflows mis en œuvre : https://cds.unistra.fr/vizier-org/OAIS_vizier_architecture.png

La certification sert à assurer la confiance des utilisateurs. Candidater à une certification a été une décision du CTS, pour sortir du contexte disciplinaire.

L'auto-évaluation permet de mesurer les marges d'amélioration. C'est un travail d'équipe extrêmement constructif. Il permet de produire des documents descriptifs des processuspublics.
La certification a conduit à décrire de bout en bout les processus en s'appuyant sur la documentation existante.

La certification a permis d'avoir des retours positifs de la part des autorités de tutelle (= reconnaissance des instances de gouvernance)

Questions :

- Avez-vous fait plusieurs certifications ? IFREMER va parler de la granularité de la certification. Il n'y pas de réponse toute faite, il faut réfléchir. Aladin a simplement été rajouté dans le dossier cette fois-ci. Il faut partir du schéma OAIS, et voir si tout peutrentrer dans la même certification. On peut donc ajouter des choses d'une certification à l'autre.
  Pour Orsay, c'est tout le jeu de données qui a été certifié à la fois, car on a montré une uniformité des règles.
  Quand la BNF candidatera-t-elle ? C'est dans sa feuille de route.

♦ Retours d'expérience CTS : IDOC-Data par Gilles Poulleau (Univ. Paris-Saclay)

IDOC collecte des données spatiales. Ça se traduit par une soixantaine de jeux de données issues d'expéditions de l'OSU Paris Saclay. Certaines de ces données sont de 2° génération, par retraitement des données primaires.
Il faut montrer que l'on a la capacité à assumer cette mission avec confiance, par exemple en intervenant rapidement sur certains instruments. La confiance est au coeur de la Science Ouverte. L'approche CTS valide certains points par des évaluateurs externes.
La certification force à repenser certaines pratiques et procédures. En particulier, à expliciter des habitudes "évidentes" par des résumés synthétiques.
Si on comptabilise tout, cela a pris énormément de temps. Mais le retour sur investissement s'obtient dans les réponses aux appels d'offre.
Attention à bien utiliser le vocabulaire approprié (OAIS, RDA, etc.).
Réponse finale acceptée à la 3° soumission.

ANO : Actions Nationales pour l'Observation
https://www.insu.cnrs.fr/fr/les-services-nationaux-dobservation
Centres et pôles de données :
https://www.insu.cnrs.fr/fr/les-centres-et-poles-de-donnees

Continuité d'accès : IDOC a renoncé à "fully implemented".

Questions :

- Comment archiver les ontologies associées, en particulier pour des installations industrielles ? Ces ontologies changent au cours du temps, et elles sont souvent distribuées, sans parler de la cybersécurité. Or il y a des obligations réglementaires -
  e.g. nucléaire. Le système d'archivage électronique doit être auto-suffisant. Il y a donc 2 solutions : aspirer les ontologies et les préserver avec les données - le choix du CINES ; soit référencer un autre entrepôt qui conserve ces ontologies.

♦ Débats ouverts

L'aspect charge de travail, mais aussi l'aspect bénéfice, étaient très intéressants et devraient en inspirer d'autres. Le CTS est aussi destiné aux bibliothèques, aux archives,

voire au secteur privé.

Merci aux intervenants. Et au revoir. Les présentations seront disponibles sur le site web du groupe PIN.

- **Prochaine réunion**

Prévue le 6 juillet 2023