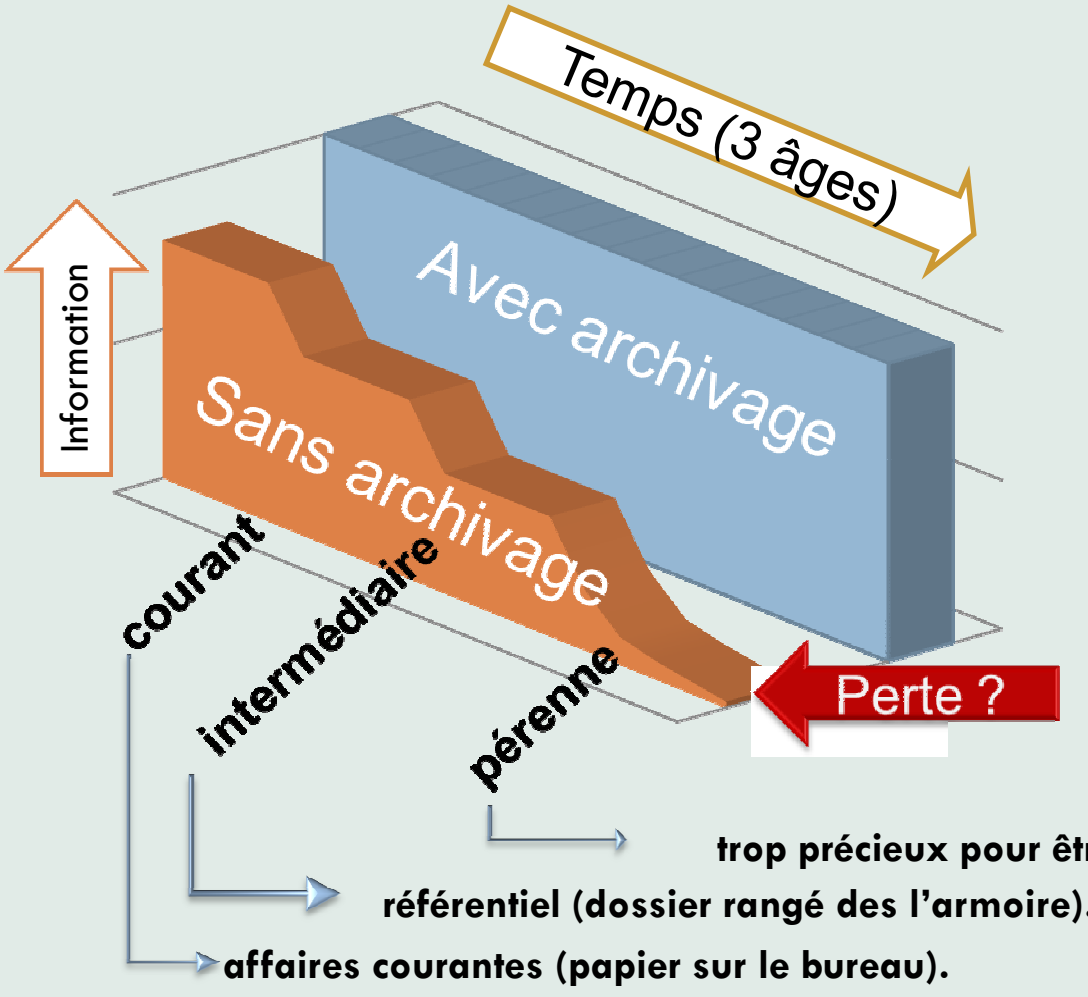
Hands-on ISAAC

Introduire des données HEP dans le projet ISAAC

Workshop PREDON

Stéphane Coutin – 14 nov 2013

Les risques liés à l'information



Risque sur :

- **Compréhension**
- **Intégrité**
- **Exploitation**
- **Valorisation**

Mise en œuvre

- Métadonnées
- Contrôle des formats
- Communauté structurée
- Contrôle d'intégrité
- Veilles

Les systèmes d'archivages du CINES

PAC

- **archivage intermédiaire et à long terme de données administratives, patrimoniales et scientifiques**
  - Mandat pour l'archivage des thèses électroniques soutenues en France (arrêté du 7 août 2006)
  - Agréé pour l'archivage intermédiaire par le SIAF
  - Périmètre opérationnel : données de l'enseignement supérieur et de la recherche
  - Partenariat avec le TGE Adonis : archivage et diffusion des données numériques en SHS

ISAAC

- **archivage intermédiaire de données scientifiques**
  - Dimensionné pour des petites structures ayant de grands volumes de données
  - Une donnée organisée et validée par des communautés d'experts
  - Un travail scientifique valorisé par le partage et la diffusion

Analyse des besoins

L'archivage des données scientifiques n'est actuellement pas pourvu de cadre fonctionnel, et organisationnel unanimement reconnu même s'il existe des initiatives dans ce sens.

Maitrise d'œuvre interne (CINES)
archivage pérenne, calcul scientifique et stockage de gros volume de données.

Prise en compte du projet EUDAT
Exigences des utilisateurs et des recommandations au niveau européen.

L'enquête ISAAC
proche des utilisateurs du CINES qui seront à priori les premiers utilisateurs.

Participations à des conférences

Analyse des besoins – Enquête projets calculs

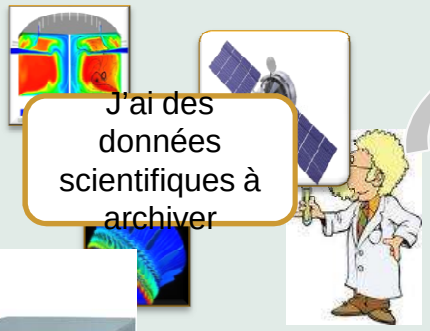
l'enquête auprès de 155 responsables de projet de calcul scientifique au CINES identifie les éléments suivants :

- **Il existe une volonté de conservation et d'accès aux données pendant au moins 3 à 5 ans,**
- **Le type de données correspond à des résultats de calcul et d'observation, code sources,**
- **Habituellement les données sont conservées dans les laboratoires au format Text, binaire, HDF5, NETCDF,FITS, Grib, CGNS.**
- **Archivage des fichiers explicatifs et des métadonnées embarquées dans les fichiers doit etre pris en compte.**
- **Il n'existe pas de jeux ou de standard de métadonnées par thématique de recherche.**
- **Nécessité d'un partage des données dans un cercle restreint et la gestion des droits sur ce partage.**
- **Grand volumes de stockages : de 1 à plusieurs dizaines de To Par Projet**

ISAAC: Description du projet

- **Service de stockage sécurisé avec un jeu minimal de métadonnées descriptives.**
- **Garantir la compréhension des données**
- **Archiver sur 3 à 5 ans puis deux possibilités :**
  - Restitution au producteur
  - Archivage pérenne
- **Travailler avec une communauté structurée d'utilisateurs**
  - Même formats, descriptions etc…

Les étapes de la vie d'un projet d'archivage dans ISAAC



J'ai des données scientifiques à archiver

Je remplis le formulaire Web de demande d'ouverture d'un projet d'archivage de données scientifiques dans ISAAC

Le Comité Thématique d'Archivage de mon domaine scientifique étudie et valide mon projet

Je travaille avec le CINES à la mise en œuvre des éléments spécifiques de mon projet.

Au terme de la durée de conservation
Je récupère mes données OU
je lance un processus d'archivage sur le long terme dans PAC (si données éligibles)

Mon projet entre en production, je réalise l'archivage de mes données sur l'interface ISAAC

Moi-même, les membres de mon équipe ou d'autres utilisateurs autorisés pouvons consulter et partager les données du projet sur le site d'ISAAC

Cas concret

- **Comment aborder un projet d'archivage sur la plateforme ISAAC**
- **Production Monte Carlo de la Collaboration H1**

Les données et méta données

- **La production Monte Carlo s'appuie sur des logiciels nommés "générateurs", l'output de ces logiciels est encore procéssé en « simulation » et ensuite « reconstruction ».**

Find Monte-Carlo Generator's File

Generator name: any
Filename: (can be part of filename)
Physics Working Group: any ID:
Lepton type: any Radiative MC: any NC/CC: any Q2 min: any
Analysis purpose:
any other:

Find Generator's File Reset form to initial values

6 records found

Rows printed: **1** - **6**

id	Generator	File name	Lumi	Events	Q2 min	Date
3891	djangoh14	/acs/mc/djangoh14/DJ14.NCHERA2.POSI.NOPOL.MRSH.Q21000.A00-A02	366.24	100002	1000	08-JAN-07
3892	djangoh14	/acs/mc/djangoh14/DJ14.NCHERA2.POSI.NOPOL.MRSH.Q210000.A00-A02	35083.05	100002	10000	08-JAN-07
3933	djangoh14	/acs/mc/djangoh14/DJANGOH14.CC.POSI.MRSH.Q210000.CDMtuned.A00-A05	15832.14	600003	100	09-FEB-07
3934	djangoh14	/acs/mc/djangoh14/DJANGOH14.CC.ELEC.MRSH.Q210000.CDMtuned.A00-A05	8519.05	600003	10000	09-FEB-07
3955	djangoh14	/acs/mc/djangoh14/DJANGOH14.CC.POSI.MRSH.Q210000.B.CDMtuned.A00-A04	721531.95	500003	10000	22-FEB-07
3956	djangoh14	/acs/mc/djangoh14/DJANGOH14.CC.ELEC.MRSH.Q210000.B.CDMtuned.A00-A04	88596.87	500003	10000	22-FEB-07

Find another generator's files

- **Proposition : l'unité archivée est la production**
  - Les métadonnées reprennent les informations de ce niveau

Generator's File

Id: 3956	Generator: **djangoh14**	Date: **22-FEB-07**	Working group: **ReX**
File name:	/acs/mc/djangoh14/DJANGOH14.CC.ELEC.MRSH.Q210000.B.CDMtuned.A00-A04		
Analysis purpose:	hadronic final state - jets		
Main cuts and comments:	CC DIS, DJANGOH + CDM QED radiative effects on, Q2 > 10000 GeV^2, no Weighting. PDF set 3036 = MRSH. Special CDM steering from HAQ di-jet CC analysis (F.Keil), tuned for High Q2.		
Lumi: **88596.87/pb**	Events: **500003**	E Energy: **27.6**	P Energy: **920**
Lepton type: **e-**	Radiative MC: **Y**	NC/CC: **CC**	Q2 Min: **>=10000**
Log file		simulated/reconstructed files	

Add new comment | Find another generator's files

  - Et pour chaque fichier la composant….

Simulated and reqonstructed files from generator's file No. 3956

2 records found

Records listed: **1** - **2**

id	p_id	Input file name	Events	Run period	Request Date	Status
4504	3956	/acs/mc/djangoh14/DJANGOH14.CC.ELEC.MRSH.Q210000.B.CDMtuned.A00-A04	500000	04_05_e-	22-FEB-07	done
4966	3956	/acs/mc/djangoh14/DJANGOH14.CC.ELEC.MRSH.Q210000.B.CDMtuned.A00-A04	500000	04_05_e-	30-OCT-07	done

dst -> mods&hat job

Niveau fichier

– L’identification et les informations liées au fichier

Request for Monte Carlo mass production

Id: 4504	Parent id: 3956
Input file name:	/acs/mc/djangoh14/DJANGOH14.CC.ELEC.MRSH.Q210000.B.CDMtuned.A00-A04
Output file name:	**/acs/mc/djangoh14/05/DJANGOH14.CC.ELEC.MRSH.Q210000.B.CDMtuned.4504.S35800.R96500.DST.A00-A30**
Working group:	ReX
Events:	500000
Run period:	04_05_e-
Beam tilt:	Y
RT noise:	N
Shower library:	N
Z vertex run simulation:	N
Steering :	
Step:	H1SIMREC
Software version:	32
H1OO Software version:	
MC Output format:	DST
Production site:	RAL
Date of request:	22-FEB-07
Status:	Done 02-MAR-07

Log file

Add new comment | Find another request | Submit MODS+HAT

- **Questions : ces informations sont elles suffisantes pour celui qui voudra comprendre et utiliser ces données dans x années?**

Autres questions à aborder

- Type d’archive
- Durée
- Qui utilisera les données (producteur, autres personnes)
- Formats des données, besoin de le valider, besoin de migrer les données
- Besoin en terme de diffusion

Démonstration ISAAC (video)

- Cas concret conçu avec le CERFACS dans le cadre du projet EUDAT
- Données de la communauté ENES (climatologie)
- Données créées à partir d'un workflow
- Archivage selon la classe de service 'compatible DSA' (Data Seal of Approval)