Ecole Thématique DATA SDUE

Guide de Survie dans la jungle des données

en Sciences de l'Univers et de l'Environnement (SDUE) :

Comment gérer les données pour les valoriser?

Session « Choix et utilisation de formats de fichiers standards, ouverts et libres »

Olivier Rouchon

Les généralités et le contexte

Donnée, information, connaissance, compétence

Crédits section : L.Duplouy(BnF) – Formation « pérennisation de l'information numérique » Aristote/PIN 2022

Que cherchons-nous à conserver ?

01010100100001101010101010001111010100001111010101010111
11101101000001010111111101010101110101010101010101011
11010101000011101010101011111101011010101011111101000010
01010101010101010111101101010101011111100001010010101010
11000011111011110101001110101010101101010101011110101010
10101011110010101001000011010101010100011110101000011110
10101011111111011010000010101111111010101011010101010101
01010101011110101010000111010101010111111010110101010111
11101000010010101010101010111101101010101111110000101
00101010101110000111111011110101001110101010101101010101
11101010101101010111100010101010111000011111011110101011

Qu’est-ce que l’on veut pérenniser ?

- Suivant que l’on veut conserver/transmettre/péreniser:
  - L’apparence visuelle d’un document dans un but de preuve et/ou de mémoire
  - Le contenu sans contrainte sur l’apparence visuelle du document qui sera restitué dans le futur
  - Des fonctionnalités complexes existant dans une base de données
- Les choix résultants
  - Choix de standards (formats de données, métadonnées)
  - Choix techniques associés

seront différents

Quelle est la problématique ?

- But = Pérennisation des informations numériques
  - C'est l'information qui doit être pérennisée
- L'information (dans ce contexte) = donnée compréhensible par un être humain
  - Vidéos, images, sons, textes.
  - Attention importance forme/contenu pour le texte et les documents composites (texte + image)
- Quelle est la différence entre une donnée, une information, une connaissance ?

Les notions

22

- Est une donnée

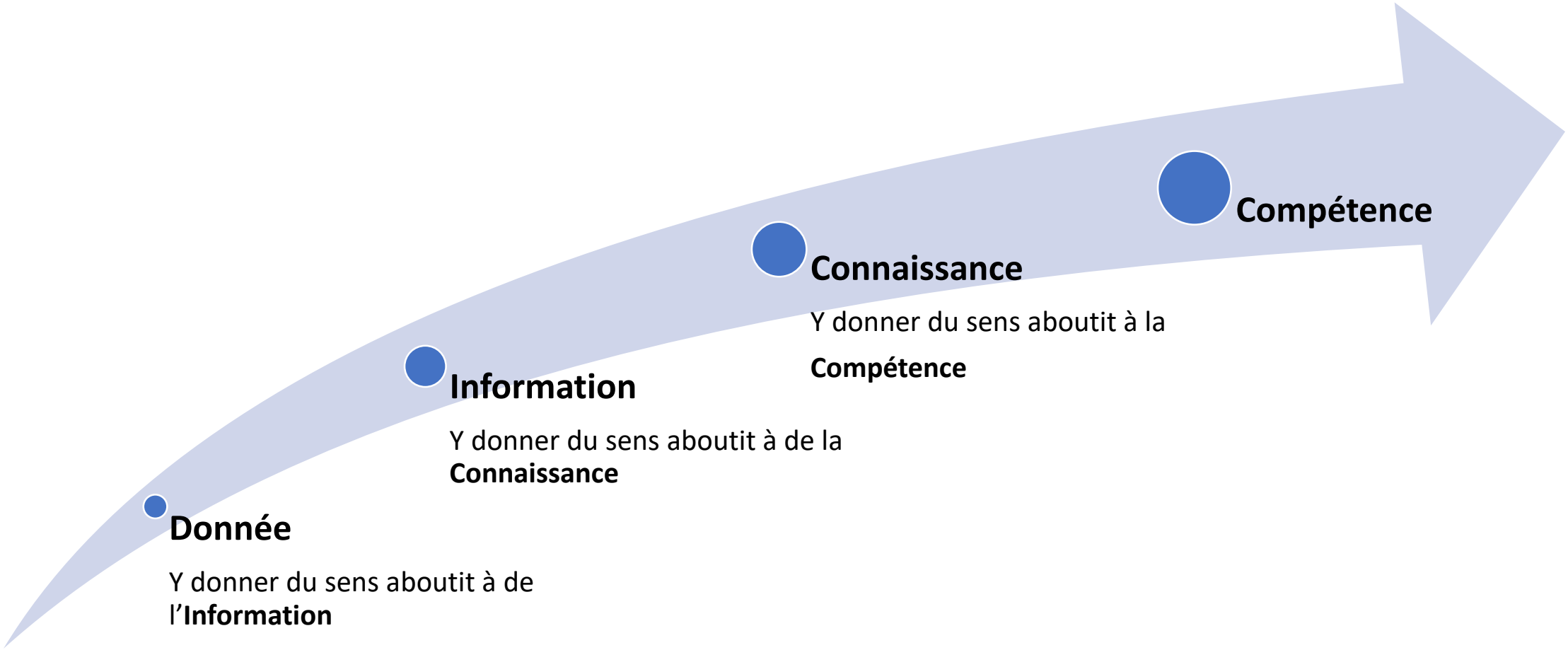
22 degrés Celsius

- Est une information

22 degrés Celsius est une température confortable

- Est une connaissance

Donnée, information, connaissance



Information vs données

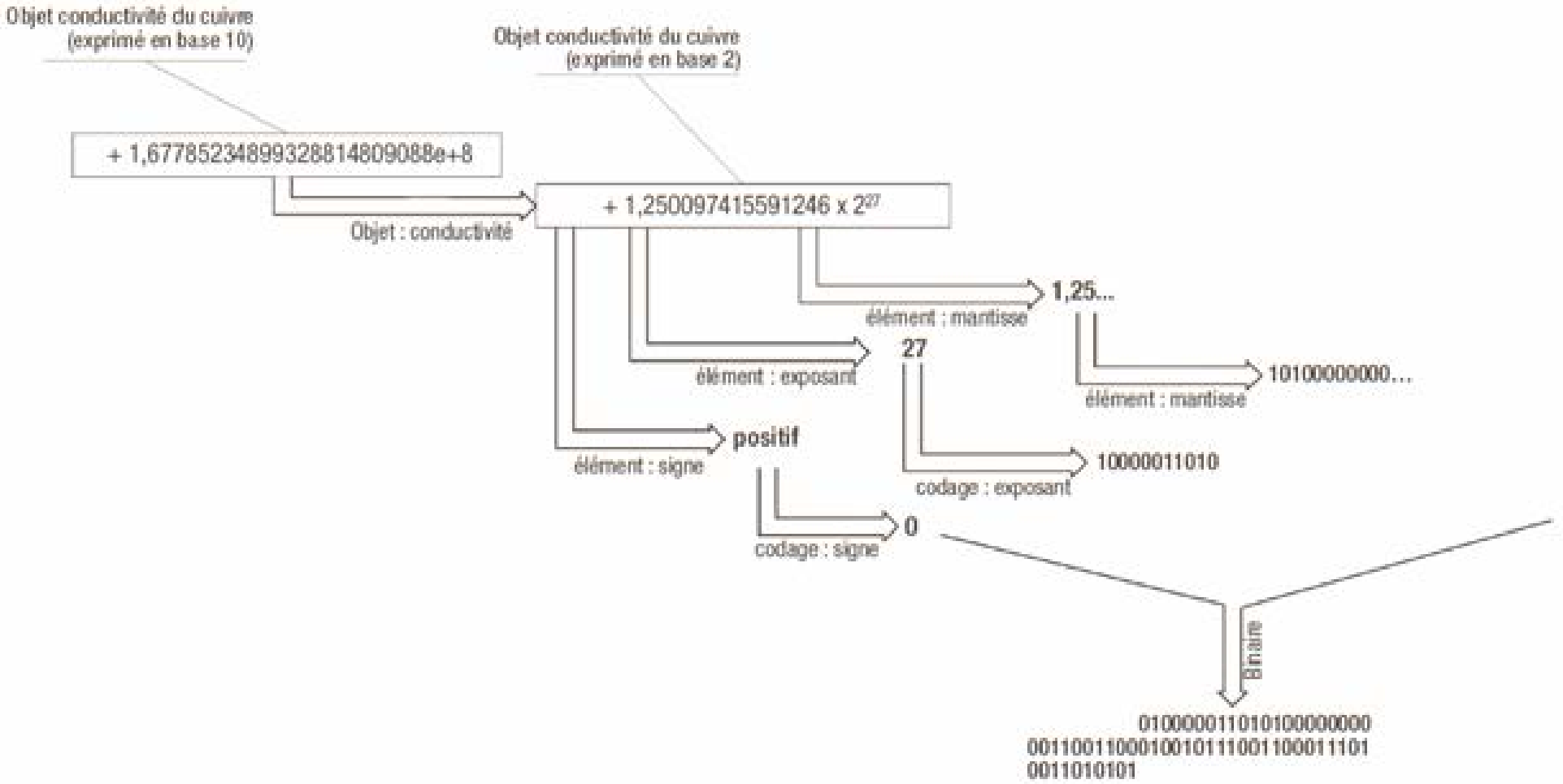
- *Information*
  - tout type de connaissance pouvant être échangée
  - indépendante des formes (physique ou numérique) utilisées pour la représenter
- *Données*
  - représentation de l'information interprétable de manière formelle permettant la communication, l'interprétation ou le traitement.

En résumé

Texte, son, vidéo = *Information*

Format de représentation = *Support d'information*

Exemple


Objet conductivité du cuivre (exprimé en base 10)
Objet conductivité du cuivre (exprimé en base 2)
+ 1,67785234899328814809088e+8
+ 1,250097415591246 x 2^27
Objet : conductivité
élément : mantisse
1,25...
élément : exposant
27
10100000000...
élément : mantisse
élément : signe
positif
10000011010
codage : exposant
codage : signe
0
Binaire
01000001101010000000
00110011000100101110011000111010101

La définition d’une politique de formats

Les critères de sélection d'un format de fichier

- Elaboré par la cellule nationale de veille sur les formats
- S'appuie sur les travaux de :
  - Library of Congress,
  - British Library,
  - Harvard Library,
  - National Archives and Records Administration (NARA)
  - Bibliothèque nationale de France (BnF).
- 9 critères retenus :
  - Adoption
  - Existence et disponibilité de la documentation
  - Transparence
  - Contenu additionnel embarqué
  - Dépendances externes
  - Incidence des brevets
  - Dispositifs techniques de protection
  - Résilience
  - Compacité

Adoption

- Utilisation du format pour produire, conserver, diffuser ou échanger l'information.
- Plus l'adoption est large
  - Plus les outils de création, de lecture, de transformation, d'émulation, etc. sont nombreux
  - Plus le format est maintenu et mis à jour.
- Pour les services patrimoniaux, l'utilisation d'un format recommandé pour la conservation par d'autres institutions est un point important de l'adoption.
- L'âge du format et sa diffusion dans plusieurs pays sont notamment des facteurs permettant de mesurer le taux d'adoption.

Existence et disponibilité de la documentation

- Spécifications disponibles pour le format :
  - Complète
  - Précise
  - Compréhensible
- Pouvant provenir d'une source institutionnelle mais pas forcément
- Existence d'outils accessibles pour identifier et valider le format.
  - Le facteur peut par exemple être corroboré par l'existence d'outils de validation élaborés par différentes sources, preuve que les spécifications sont accessibles et complètes.
  - Les formats ouverts présentent généralement une mise à disposition de la documentation plus complète que les formats propriétaires mais ce n'est pas forcément systématique.

Transparence

- Accessibilité du format à l'aide d'outils non spécialisés
  - Possibilité pour un humain de lire et de comprendre le code du format grâce à des outils de type éditeur de texte.
  - Outils qui servent à produire le format doivent par ailleurs être utilisables sur des machines standard.
- Suppose également que le format soit non compilé et encodé avec des caractères standardisés.
- La compression, si elle est inévitable, doit utiliser des algorithmes largement répandus ou précisément documentés par l'utilisateur.

Contenu additionnel embarqué

- Capacité d'un format à être autoporteur
  - Comprendre au sein de son code les flux de données nécessaires à sa compréhension et son exploitation.
- Ces flux sont de plusieurs types :
  - Représentation (permet de transformer des données en information exploitable),
  - Identification,
  - Contexte (lien entre une information et son environnement),
  - Intégrité (traçabilité des modifications),
  - Provenance (historique de l'information)
- Inverse d'un format qui obligerait l'utilisateur à stocker ces flux séparément de l'information primaire.

Dépendances externes

- Nécessité d’utiliser
  - Des logiciels (notamment propriétaires)
  - Du matériel informatique spécifiques (joystick, microphone, etc.) pour exploiter le format.
- Ce facteur concerne particulièrement les contenus dynamiques, comme les jeux vidéo ou certaines œuvres d’art numérique

Incidence des brevets

- Présence de brevets ou de toute contrainte juridique sur un format
  - Peut entraîner des coûts de conservation importants et limiter le développement d'outils d'analyse et d'exploitation du format.
- Présence d'un brevet (ou d'une réglementation similaire) doit être associée à l'analyse de l'adoption du format
  - Bon indicateur pour savoir si le brevet empêche la large diffusion du format et la création d'outils d'exploitation.

Dispositifs techniques de protection

- Manipulation de l'information par les services de conservation à des fins de
  - Diffusion,
  - Transformation,
  - Réparation, etc.
- Empêchée par
  - Le chiffrement,
  - La protection par mot de passe,
  - Eventuellement la présence d'informations de copyright sous forme de watermark,
- Attention d'autres dispositifs comme la signature électronique, n'empêchent pas la manipulation du contenu, mais sont conçus pour assurer son authenticité.

Résilience / Compacité

- Résilience
  - Capacité du format à permettre la lisibilité de son contenu, même en cas d'altération partielle de celui-ci.
- Compacité
  - Capacité du format à contenir un grand nombre de données dans un volume réduit.
    - Si la compacité est liée à la compression, s'assurer de la transparence et de la réversibilité de la compression.

- Format maîtrisé = identifié et vérifiable
  - Format publié ; ex. WAVE, SVG ;
  - Format largement utilisé ; ex. XML, MPEG4 ;
  - Format normalisé si possible ; ex. PDF (ISO 32000-1:2008), PNG (ISO 15948:2004).
- Respect des spécifications du format
- Existence d'outils libres permettant une identification, validation et caractérisation.

Type	Format
Texte	TXT, XML, HTML, PDF, ODT
Image	GIF, JPEG, TIFF, PNG, SVG, JPEG200
Son	WAV, AIFF, AAC, OGG (VORBIS)
Vidéo	MPEG4, OGG (THEORA), MKV
Données brutes	CSV, XML, ODS, HDF5, NetCDF, FITS

Quelques exemples

PDF, NetCDF, FITS

PDF - Portable Data Format

- Langage de description de page présenté par la société Adobe Systems en 1992
- Norme ISO depuis 2008
- Préservation des
  - polices de caractères, images, objets graphiques,
  - mise en forme

de tout document source, quelles que soient le logiciel de lecture

- Format utilisé dans de nombreux logiciels
  - Exportation dans les suites bureautiques grand public,
  - Manipulations par des programmes spécialisés
  - Génération de documents officiels
- Format binaire

PDF - Portable Data Format

- Spécifications : https://www.iso.org/fr/standard/51502.html (v1.7, 2008)
- PUID : fmt/14 (https://www.nationalarchives.gov.uk/pronom/fmt/14), fmt/276
- Type MIME : application/pdf, application/x-pdf, application/x-bzpdf, application/x-gzpdf
- Signature : 25 50 44 46

PDF - Portable Data Format

%PDF-1.5
%µµµµ
1 0 obj
<</Type/Catalog/Pages 2 0 R/Lang(fr-FR) /StructTreeRoot 106 0 R/MarkInfo<</Marked true>>>>
endobj
2 0 obj
<</Type/Pages/Count 11/Kids[3 0 R 28 0 R 48 0 R 68 0 R 83 0 R 90 0 R 92 0 R 97 0 R 99 0 R 101 0 R 103 0 R] >>
endobj
3 0 obj
<</Type/Page/Parent 2 0 R/Resources<</Font<</F1 5 0 R/F2 12 0 R/F3 14 0 R/F4 16 0 R/F5 21 0 R/F6 23 0 R>>/XObject<</Image7 7 0 R/Image9 9 0 R/Image11 11 0 R>>/ProcSet[/PDF/Text/ImageB/ImageC/ImageI] >>/MediaBox[0 0 594.96 842.04]
/Contents 4 0 R/Group<</Type/Group/S/Transparency/CS/DeviceRGB>>/Tabs/S/StructParents 0>>
endobj
4 0 obj
<</Filter/FlateDecode/Length 4764>>
stream
xœÕ][¹⁄₉~7àÿPŘUïU
£±ÝÍx±Î¶Çíl OdYöôÀ-õHê óoóø7äeÏ!YW^ªº ™¸»UEò|<$Ï‡ÔÂ‡Óíçõæ”<{vñÃé´Þü²ý”|¸x¾?ðw□»¸ùý~{q½þr»[Ÿn÷»‹wOøèÕ~□Ú./"ç/_$Ïož>¹xEBÒŒ'7ŸŸ>!Iÿ'DÑ4£<aE?oîž>É'/øãOOŸ|X$Ë¿%7ÿýôÉTÿíé“·ú§àYšX£□
Ã$Kóä°}úäÿþ³2$R¤Ã“wá˜Šœ¦<×h>,HÃÛ©ð4“Ö§ô·úÿÉÕ›IrqL□óâõË$³œ„$§)ºMh0ÉŠ'"‰Dw9O6Ðð Åë»õ—J^î_c¤nLQÝüâØl"sŽ°ëFŠP#´3¶Nws™'œ'ª)5ý5o_Ê¿~rFÙ¡ÂýÎrÓ2£©IÁ°ÓR^T4ñÌè¦¤L%+ëÏÜ4Ëg¤žIK3à4)ë›Y›ó²”ëÞÕŸÜšQn `-
¢ ì”§T(ª4Kq+4{+.þg½û',¶»ÕŸž/ë®7[â¦¥GuP\}J©”É²ëðÜvþuÙ7$ò¼ %,Oùb¹b‹ë¥X\ýäÅ}„¢M–%,Ð^ò.¦õ›Ã¨¨>ÂÉ(/ êæÃ,fÔ•þSa¨<½0>,V³ÎÊÃGö_Í
ƒ''ý¶FÔ™Ö^ ,%J@½|2Èü¼
™©”J«2¯Þ¾□}½ä‹«7KBW□^æ‹›ws._–¹D½\˜¯□æ-µ¤Þ-™L¯¡oÐÃ›×¯^¿]¶x¿\ÑÅU§«ÎH³1Ÿ]S¨6¸a³Ð³OS,õ9Kú“Q…¶Q¥?ahãê“4l$]¹6êÊùKŒ²~•ùLêÍa¿[®øâ÷;ñm²”‹O0ðYhN…´Õ=à8□ÝžÔÅÒ‡g"NŠÌ·™ä&¡³□
ç…6°ç'•„Í¬U“€ñQÅ˜oRAQa[`b±?Ü?Àt:&?.W¦:Lü»õîánÊôòàâTbõ^Lƒ}2",¿°®ñkÐ(`@ò}ð ò,#ú·†øù‹èŽÚ"¦£\pôê6TQä:Þ'X~–Híßµ □?¶T5€P²ÎW³ŒÅ]ƒÒ¸; …¹ªs¦"¥Üa•
Ï¶§\¦²ëlö¸°DyÚ³Î±d"-|Á ·•<Òè@Z©ÜMj”ô„”æ4•$¨þŠ¶³]wkS¯-…‘V„‡ Ç #XUELxÞ Ž»yp:l—«ú!Ÿ@ñ'LÿyX®òÅþWP>Û<¥úéV¿ßùìé E'¥9‹€þ| ¸^®äb÷Oø¥tŽˆtvHh:'‚„>A"ÓL” åø):
hÓ□
ÕöÃo[q”¤” vU‹Þ:Ì÷¤ÿ~©ø}L|0L”Ð˜?,*P‹õàW1LÆ@Q0,`f‹G¤Ò_ô 44)ÅÑÕâ¶%s¼GÔö=áyZðv Þ%x×Á,3ì{ËŒV À¬p5C□ ÛËrýªÁðÎH C¥. ©tÊt$“hVÔk‡ç¡™K]lf‹
'º¯»gYÆÈ%‡_,Â¿þÇ.W9|æÞå—«â™~&ðÑ+ÿë¹¹¼\a
þâ2$,`ôÊŠ<e½□_°ÜÖ¸ÜNZ'¢8{ØœÐÓ|¨lÃ²l¼'¥<KÁ·Ùï6e«aî¾mñ×Ù`Ð´€y×
&éq¼)5ã"ìáçüIoè×1¼\¸+è£•
xô`x#çÊy0öÑÆ?Ü› &Áè‡?,ÜÇwk×wKi>åüdÔ'EÙ1_û
É—©JþjÙÕâ¶H-Ä[…ò1ÔNC¯A'Z"·ÕRÔ4SÊê."
Úp®Æ®V °dª÷¥°niËj^0½I1@VKgåÄ+÷®à›Š ¾w·0åNÛ$Iw¤†PJ°±tešËtï÷(ÍQÐƒL□@?k;ž'Å.Æ{ý¾Øë¿~E+û43bRpŒ{>–Sî¢TsÎœ\ÆGhWæZúCdnîÈÜxå¸Ì;–^>É6•×¶²ŸÔûL•[TÑ□7ž[Ž·hµŸPè□Ã8Ù®×VÞ□[ÆÒ¼aól5M!%00
õèÝú+v[÷øðûRên¿1KJhý¦9rï5Ÿ¾"×¾–ì3¬ÃÄÈJžý¼x□ýxðÍ~`¼•ûñórVZ”¡•?l–z—zÑÑ¿ ‰©ýŸfH§a~Òt¢š…/ÔNÃÞ ¤éJ5
ÕŠÑ4ã
2YÐ–s•3Ô|oóRª×•«R¾"ÃH[#ÈB,s%h¬nŸ åR¥ŠDèOÛã}¥\vÇÒÿø¨?□B:'è¤ÖH£ÐpÿsÒ¾_üòDˆñÄ•þ±ºQáÏU^{zõ²zuXï–ÅâŸ{ZVk{û:¸°G Ž`$™kgM —=P~zëD
ÛÉáŠ
v¯Í<atÎ1–9Zø]×‡R¯|.µîñ¸Å
•ƒÙ=,uŒzT»¨□zÔôqe×õâ,ÍX ?FAˆýÏ,ÂLÊ ³æœL¨Ë]Üƒf©æ1r¾3"Z°ipëñûwûõ

NetCDF - Network Common Data Form

- Format de fichiers pour données scientifiques stockées sous la forme de tableaux de nombres multi-dimensionnels
- Pluridisciplinaire, très utilisé en météorologie, océanographie
- Format binaire
- Structure de données « auto-documentée »
  - En-tête
    - Décrivant la disposition des données dans le reste du fichier,
    - Contenant une liste arbitraire de métadonnées présentées sous la forme d'attribut de type nom/valeur.
  - Tableaux de données
- Indépendant de l'architecture matérielle

NetCDF - Network Common Data Form

- Spécifications : https://docs.unidata.ucar.edu/netcdf-c/current/file_format_specifications.html (v4.0, 2008)
- PUID : fmt/282 (https://www.nationalarchives.gov.uk/pronom/fmt/282), fmt-283
- Type MIME : application/netcdf, application/x-netcdf
- Signature : 43 44 46 01
- Bibliothèques fournies par l'UCAR : https://www.unidata.ucar.edu/software/netcdf/
- File format validation : https://www.seanoe.org/data/00344/45538/

NetCDF - Network Common Data Form

```
CDF          lon   ´   lat   ª   time       bnds
   title        IIPSL  model output prepared for IPCC Fourth Assessment SRES A2 experiment
institution       3IPSL (Institut Pierre Simon Laplace, Paris, France)    source       ⍰IPSL-CM4_v1 (2003) : atmosphere : LMDZ (IPSL-CM4_IPCC, 96x71x19) ; ocean ORCA2 (ipsl_cm4_v1_8, 2x2L31); sea ice LIM (ipsl_cm4_v    contact       2Sebastien Denvil, sebastien.denvil@ipsl.jussieu.fr
project_id       IPCC Fourth Assessment     table_id      Table O1 (13 November 2004)
experiment_id        SRES A2 experiment
realization            cmor_version      ?uÂ⍰
Conventions      CF-1.0     history      «YYYY/MM/JJ: data generated; YYYY/MM/JJ+1 data transformed  At 16:37:23 on 01/11/2005, CMOR rewrote data to comply with CF standards and IPCC Fourth Assessment requirements    references       7Dufresne et al, Journal of Climate, 2015, vol
XX, p 136    comment       Test drive
     lon
standard_name             longitude                      long_name                   longitude     units       degrees_east   axis     X     bounds      lon_bnds   original_units       @degrees_east                                          lon_bnds
@  °   lat
standard_name       latitude                             long_name       latitude   units
degrees_north     axis     Y     bounds      lat_bnds   original_units       @degrees_north                                    P  ð   lat_bnds                   @   time
standard_name        time                                long_name       time   units       =days since 2001-1-1                                   axis     T     calendar     360_day    bounds                                   time_bnds     original_units       seconds since 2001-1-1        *à
                              time_bnds                    *è   tos
standard_name       sea_surface_temperature     long_name       Sea Surface Temperature    units       K      cell_methods      !time: mean (interval: 30 minutes)     _FillValue       `xì
missing_value       `xì
original_name       sosstsst   original_units      degC      history     ¯ At   16:37:23 on 01/11/2005: CMOR altered the data in the following ways: added 2.73150E+02 to yield output units;  Cyclical dimension was output starting at a different lon;    Þ   *ø?ð     @     @     @     @"     @&
@*    @.    @1    @3    @5    @7    @9    @;    @=    @?    @@€    @A€    @B€    @C€    @D€    @E€    @F€    @G€    @H€    @I€    @J€    @K€    @L€    @M€    @N€    @O€    @P@    @PÀ    @Q@    @QÀ    @R@    @RÀ    @S@    @SÀ    @T@    @TÀ    @U@    @UÀ
@V@    @VÀ    @W@    @WÀ    @X@    @XÀ    @Y@    @YÀ    @Z@    @ZÀ    @[@    @[À    @\@    @\À    @]@    @]À    @^@    @^À    @_@    @_À    @`    @``    @`    @`à    @a    @a`    @a    @aà    @b    @b`    @b    @bà    @c    @c`    @c    @cà    @d    @d`
@d    @dà    @e    @e`    @e    @eà    @f    @f`    @f    @fà    @g    @g`    @g    @gà    @h    @h`    @h    @hà    @i    @i`    @i    @ià    @j    @j`    @j    @jà    @k    @k`    @k    @kà    @l    @l`    @l    @là    @m    @m`    @m    @mà    @n    @n`    @kÀ
@kÀ    @l    @l    @l@    @l@    @l€    @l€    @lÀ    @lÀ    @m    @m    @m@    @m@    @m€    @m€    @mÀ    @mÀ    @n    @n    @n@    @n@    @n€    @n€    @nÀ    @nÀ    @o    @o    @o@    @o@    @o€    @o€    @oÀ    @oÀ    @p    @p    @p    @p
@p@    @p@    @p`    @p`    @p€    @p€    @p    @p    @pÀ    @pÀ    @pà    @pà    @q    @q    @q    @q    @q@    @q@    @q`    @q`    @q€    @q€    @q    @q    @qÀ    @qÀ    @qà    @qà    @r    @r    @r    @r    @r@    @r@    @r`    @r`    @r€    @r€
@r    @r    @rÀ    @rÀ    @rà    @rà    @s    @s    @s    @s    @s@    @s@    @s`    @s`    @s€    @s€    @s    @s    @sÀ    @sÀ    @sà    @sà    @t    @t    @t    @t    @t@    @t@    @t`    @t`    @t€    @t€    @t    @t    @tÀ    @tÀ    @tà    @tà    @u    @u    @u
@u    @u@    @u@    @u`    @u`    @u€    @u€    @u    @u    @uÀ    @uÀ    @uà    @uà    @v    @v    @v    @v    @v@    @v@    @v`    @v`    @v€    ÀSà    ÀS    ÀS`    ÀS    ÀRà    ÀR    ÀR`    ÀR    ÀQà    ÀQ    ÀQ`    ÀQ    ÀPà    ÀP    ÀP`    ÀP    ÀOÀ    ÀO@    ÀNÀ
xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xìC‰îŽ`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`-
xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`-
xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xìC‰Õ6C‰î CŠaCŠ?CŠECŠVCŠC‰ö_C‰ä¾C‰×ÏC‰Ç>C‰°rC‰–ýC‰sCC‰HÏC‰
ÔCˆ±êCˆQCˆÆC‡®n`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xìCˆ;êC‡–-C‡–-C‡–-C‡–:C‡–:C‡Ç⍰ˆÊCˆ3YCˆH©Cˆ?2Cˆ$
C‡ÿC‡Ò¿C‡©4`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xìC‰ÊÇC‰ÆIC‰ÏLC‰â
C‰í¾C‰ñÚC‰ìöC‰àŠC‰ÍFC‰»’C‰¬õC‰˜UC‰x"C‰MGC‰(Cˆ™†CˆÀC‡ºtC‡¤C‡©‹C‡¦ÆC‡¥>C‡£åC‡¢¤C‡¢^C‡¢⍰xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xì`xìCˆ;êCˆ;êC‡˜‰C‡˜‰C‡–-C‡–
```

FITS - Flexible Image Transport System

- Format de fichiers pour la communauté Astronomie
  - Conçu spécifiquement pour des données scientifiques
  - Inclut beaucoup de dispositifs pour décrire l'information photométrique et spatiale de calibrage, ainsi que les métadonnées de l'origine de l'image.
- Structure multiple
  - Successions de HDU (« Header » + « Data Unit »)
  - Les métadonnées de l'image ou du tableau multidimensionnel sont stockées dans un en-tête lisible par un humain, au format ASCII
  - 3 modèles de données
    - Image,
    - Table ASCII
    - Table binaire

FITS - Flexible Image Transport System

- Spécifications : http://fits.gsfc.nasa.gov/standard30/fits_standard30.pdf (v3.0 2008)
- PUID : x-fmt/383 (https://www.nationalarchives.gov.uk/pronom/x-fmt/383)
- Type MIME : application/fits, image/fits
- Signature : 53 49 4D 50 4C 45 20 20 3D 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 54
- Support via des bibliothèques standard disponible pour la plupart des langages : C, Fortran, Java, Perl, Python, et IDL.
- « FITS Support Office » : https://fits.gsfc.nasa.gov/ - liste des bibliothèques et plateformes supportant le format FITS.
- FITS file verifier : https://fits.gsfc.nasa.gov/fits_verify.html

FITS - Flexible Image Transport System

SIMPLE = T / file does conform to FITS standard BITPIX = -32 / number of bits per data pixel NAXIS = 3 / number of data axes NAXIS1 = 200 / length of data axis 1 NAXIS2 = 200 / length of data axis 2 NAXIS3 = 4 / length of data axis 3 EXTEND = T / FITS dataset may contain extensions COMMENT FITS (Flexible Image Transport System) format is defined in 'AstronomyCOMMENT and Astrophysics', volume 376, page 359; bibcode: 2001A&A...376..359H BSCALE = 1.0E0 / REAL = TAPE*BSCALE + BZERO BZERO = 0.0E0 / OPSIZE = 2112 / PSIZE of original image ORIGIN = 'STScI-STSDAS' / Fitsio version 21-Feb-1996 FITSDATE= '2004-01-09' / Date FITS file was created FILENAME= 'u5780205r_cvt.c0h' / Original filename ALLG-MAX= 3.777701E3 / Data max in all groups ALLG-MIN= -7.319537E1 / Data min in all groups ODATTYPE= 'FLOATING' / Original datatype: Single precision real SDASMGNU= 4 / Number of groups in original image CRVAL1 = 182.6311886308 CRVAL2 = 39.39633673411 CRPIX1 = 420. CRPIX2 = 424.5 CD1_1 = -1.067040E-6 CD1_2 = -1.259580E-5 CD2_1 = -1.260160E-5 CD2_2 = 1.066550E-6 DATAMIN = -7.319537E1 / DATA MIN DATAMAX = 3.777701E3 / DATA MAX MIR_REVR= T ORIENTAT= -85.16 FILLCNT = 0 ERRCNT = 0 FPKTTIME= 51229.798574 LPKTTIME= 51229.798742 CTYPE1 = 'RA---TAN' CTYPE2 = 'DEC--TAN' DETECTOR= 1 DEZERO = 316.6452 BIASEVEN= 316.6715 BIASODD = 316.6189 GOODMIN = -5.064006 GOODMAX = 2552.17 DATAMEAN= 0.4182382 GPIXELS = 632387 SOFTERRS= 0 CALIBDEF= 1466 STATICD = 0 ATODSAT = 16 DATALOST= 0 BADPIXEL= 0 OVERLAP = 0 PHOTMODE= 'WFPC2,1,A2D7,LRF#4877.0,,CAL' PHOTFLAM= 3.447460E-16 PHOTZPT = -21.1 PHOTPLAM= 4884.258 PHOTBW = 20.20996 MEDIAN = -0.175651 MEDSHADO= -0.121681 HISTWIDE= 1.033711 SKEWNESS= -1.983727 MEANC10 = 0.12958 MEANC25 = 0.3129676 MEANC50 = 0.4577668 MEANC100= 0.3916293 MEANC200= 0.3115222 MEANC300= 0.3295493 BACKGRND= -0.3676353 ORIGIN = 'NOAO-IRAF FITS Image Kernel December 2001' / FITS file originator DATE = '2004-01-09T03:26:36' IRAF-TLM= '03:26:36 (09/01/2004)' FILETYPE= 'SCI ' / type of data found in data file TELESCOP= 'HST' / telescope used to acquire data INSTRUME= 'WFPC2 ' / identifier for instrument used to acquire data EQUINOX = 2000.0 / equinox of celestial coord. system / WFPC-II DATA DESCRIPTOR KEYWORDS ROOTNAME= 'u5780205r' / rootname of the observation set PROCTIME= 5.3013140196 76E+04 / Pipeline processing time (MJD) OPUS_VER= 'OPUS 14.5a ' / OPUS software system version number CAL_VER = ' ' / CALWP2 code version / SCIENCE INSTRUMENT CONFIGURATION MODE = 'FULL' / instr. mode: FULL (full res.), AREA (area int.)SERIALS = 'OFF' / serial clocks: ON, OFF / IMAGE TYPE CHARACTERISTICS IMAGETYP= 'EXT ' / DARK/BIAS/IFLAT/UFLAT/VFLAT/KSPOT/EXT/ECAL CDBSFILE= 'NO ' / GENERIC/BIAS/DARK/PREF/FLAT/MASK/ATOD/NO PKTFMT = 96 / packet format code / FILTER CONFIGURATION FILTNAM1= 'FR533P15' / first filter name FILTNAM2= ' ' / second filter name FILTER1 = 69 / first filter number (0-48) FILTER2 = 0 / second filter number (0-48) FILTROT = 15.0 / partial filter rotation angle (degrees) LRFWAVE = 4877.000000 / linear ramp filter wavelength / INSTRUMENT STATUS USED IN DATA PROCESSING UCH1CJTM= -88.2569 / TEC cold junction #1 temperature (Celsius) UCH2CJTM= -88.6697 / TEC cold junction #2 temperature (Celsius) UCH3CJTM= -88.3028 / TEC cold junction #3 temperature (Celsius) UCH4CJTM= -88.7671 / TEC cold junction #4 temperature (Celsius) UBAY3TMP= 13.2302 / bay 3 A1 temperature (deg C) KSPOTS = 'OFF' / Status of Kelsall spot lamps: ON, OFF SHUTTER = 'A' / Shutter in place at beginning of the exposure ATODGAIN= 7.0 / Analog to Digital Gain (Electrons/DN) / RSDP CONTROL KEYWORDS MASKCORR= 'COMPLETE' / Do mask correction: PERFORM, OMIT, COMPLETE ATODCORR= 'COMPLETE' / Do A-to-D correction: PERFORM, OMIT, COMPLETE BLEVCORR= 'COMPLETE' / Do bias level correction BIASCORR= 'COMPLETE' / Do bias correction: PERFORM, OMIT, COMPLETE DARKCORR= 'COMPLETE' / Do dark correction: PERFORM, OMIT, COMPLETE FLATCORR= 'SKIPPED ' / Do flat field correction SHADCORR= 'OMIT ' / Do shaded shutter correction DOSATMAP= 'OMIT ' / Output saturated pixel map DOPHOTOM= 'COMPLETE' / Fill photometry keywords DOHISTOS= 'OMIT ' / Make histograms: PERFORM, OMIT, COMPLETE OUTDTYPE= 'REAL ' / Output image datatype: REAL, LONG, SHORT LINENUM = '02.030 ' / PEP proposal line number SEQLINE = ' ' / PEP line number of defined sequence SEQNAME = ' ' / PEP define/use sequence name HISTORY MASKFILE=uref$f8213081u.r0h MASKCORR=COMPLETED HISTORY PEDIGREE=INFLIGHT 01/01/1994 - 15/05/1995 HISTORY DESCRIP=STATIC MASK - INCLUDES CHARGE TRANSFER TRAPS HISTORY BIASFILE=uref$j9a1612mu.r2h BIASCORR=COMPLETED HISTORY PEDIGREE=INFLIGHT 29/08/98 - 21/08/99 HISTORY DESCRIP=not significantly different from j6e16008u. HISTORY DARKFILE=uref$j2g1549cu.r3h DARKCORR=COMPLETED HISTORY PEDIGREE=INFLIGHT 16/02/1999 - 16/02/1999 HISTORY DESCRIP=Pipeline dark: 120 frame superdark with hotpixels from HISTORY 16/02/99 HISTORY FLATFILE=uref$f4i1559cu.r4h FLATCORR=SKIPPED HISTORY PEDIGREE=DUMMY 18/04/1995 HISTORY DESCRIP=All pixels set to value of 1. Not flat-fielded. HISTORY PC1: bias jump level ~0.100 DN. HISTORY The following throughput tables were used: HISTORY crotacomp$hst_ota_007_syn.fits, crwfpc2comp$wfpc2_optics_006_syn.fits,HISTORY crwfpc2comp$wfpc2_lrf_004_syn.fits[wave#], HISTORY crwfpc2comp$wfpc2_dqepc1_005_syn.fits, HISTORY crwfpc2comp$wfpc2_a2d7pc1_004_syn.fits, HISTORY crwfpc2comp$wfpc2_flatpc1_003_syn.fits HISTORY The following throughput tables were used: HISTORY crotacomp$hst_ota_007_syn.fits, crwfpc2comp$wfpc2_optics_006_syn.fits,HISTORY crwfpc2comp$wfpc2_lrf_004_syn.fits[wave#], HISTORY crwfpc2comp$wfpc2_dqewfc2_005_syn.fits, HISTORY crwfpc2comp$wfpc2_a2d7wf2_004_syn.fits, HISTORY crwfpc2comp$wfpc2_flatwf2_003_syn.fits CTYPE3 = 'GROUP_NUMBER' / Extra dimension axis name CD3_3 = 1 / CD1_3 = 0 / CD2_3 = 0 / CD3_2 = 0 / CD3_1 = 0 / END

¿Å«”?j»þ½Ã›ª?]WO?€‡›)?[Ù®?`Î ¿%‰É’¿nÒ]?OaÀ¿ƒ¬ø?ý˜à¿Š×Ü?WH’?Yëc?LÌ?C>¾?XÇP½º‹H?rsÄ¿ƒí±¿†j¿,t#¾ Kš¾!yû?ƒ1¼Ñw₧Aç?[þ<¾Atë?rù0?@À˜¾îx?c,º¾°¾ ½·(þ?NXç?ú/¾sô{¾lü€?/®`¾!]?Äi¾ãò?!E¸½ÿœ½û.`½iÃ¿•º é €?Û?öDô?3á?N¾&¿ˆÒÊ½!n¼¾ `?Pcõ½º;®¾ ^?Jù€?[2?Yüð½Â³ ¾ ‡¡?gýÃ?RÉ½ÀÊ¿×…U?]+=¾Þ¿“w"¿₧¿’òš¾5½v¾“`ò¾Vjl¼1¯ª½¹ŠT¿÷×v?è¼´¾šÆ?UD¾%¾ò¾9 Æ¿–#t¿§o!?O«À êŠ?K*L½íû8?•t@ ŒÎ?åÑ¿ª^:¿«ÎD¿ƒN©?6|p¾cF²½OóX¿ŠsÀ¿… Õ¿’?;O–?_«²?a+3¿₧-?!Û=¾|(š¾úü¾eËÞ?\0ä¿Žò{?@³Æ?+R×¿₪¼m”2?bÍk¾7J©¾Ÿ=Eh?W>²¾–Žº?ï¶-¾$¯n?G>ÿ?@ÆZ¿™èî>¢L¼Ãm’¾e½°.Ü¿•ªæ?TX›¾%¸J¿,f2?q”z?R>”¿₪¾

L’assurance qualité sur les formats

Les contrôles, les outils

Crédits section : L.Duplouy(BnF) – Formation « pérennisation de l’information numérique » Aristote/PIN 2022

Les types de contrôle possible

- Identification
  - Déterminer le type de format présumé de l'objet numérique par signature
- Validation
  - Déterminer la conformité du format par rapport à sa spécification (standard ou norme) ou par rapport à un profil d'application du format (contexte spécifique/local-use)
- Caractérisation
  - Déterminer les propriétés (nécessite des outils) de l'objet numérique
- Évaluation
  - Déterminer le niveau de risque du format (peut varier selon propriétés)

Les outils

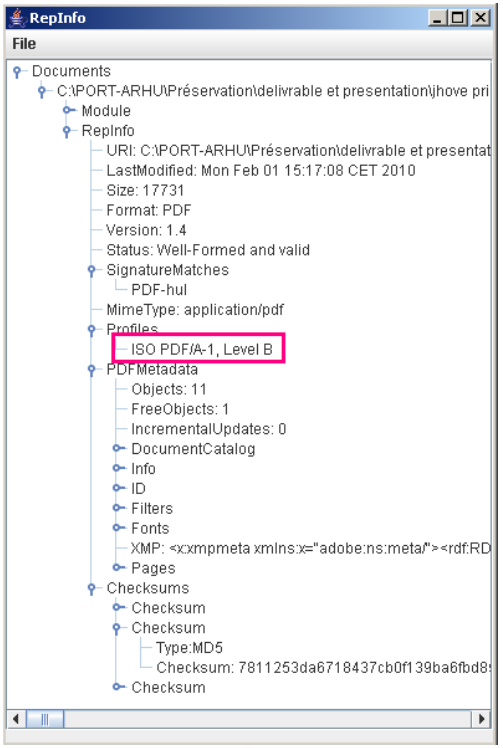
- Commande « file » sous *nix
  - Basé sur la signature
  - Paramétrable
  - Ne permet que l'identification
- DROID
  - http://www.nationalarchives.gov.uk/aboutapps/pronom/droid.htm
  - identification automatique des formats
  - utilise les signatures internes et externes référencées dans la base Pronom (PUID, ex :fmt/40)
  - beaucoup de formats sont justes signalés et pas décrits
  - les spécifications sont externes (= problème pour la préservation, pas d'évaluation possible)

Les outils

- FIDO (Format Identification for Digital Objects)
  - porté par openplanets foundation (financé par le projet Eu SCAPE)
  - http://www.openplanetsfoundation.org/software/fido/
  - Outil d'identification de format
  - Fournit des identifiants pronom
  - Va plus loin que DROID : analyse complète des fichiers
    - Pas uniquement l'entête de fichier
    - Détecte les formats imbriqués

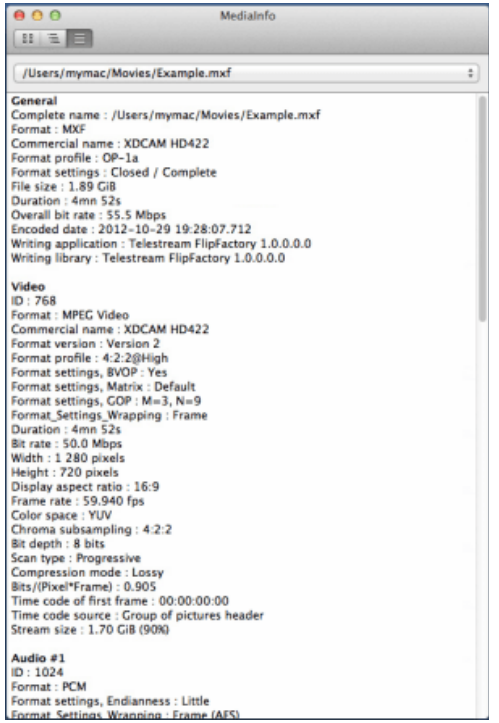
Les outils

- JHOVE
  - Initialement porté par l'université d'Harvard
  - http://jhove.sourceforge.net/
  - Outil de caractérisation
  - Fournit une description complète des formats supportés
  - Très populaire notamment pour l'analyse de fichier image (TIFF, JPEG, …) remplacé par JHOVE2
    - https://bitbucket.org/jhove2/main/wiki/Home
    - Porté par la Library of Congress
  - Amélioration du framework pour faciliter l'intégration de nouveaux modules



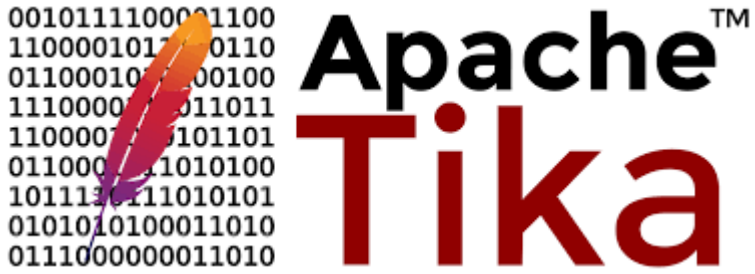
Les outils

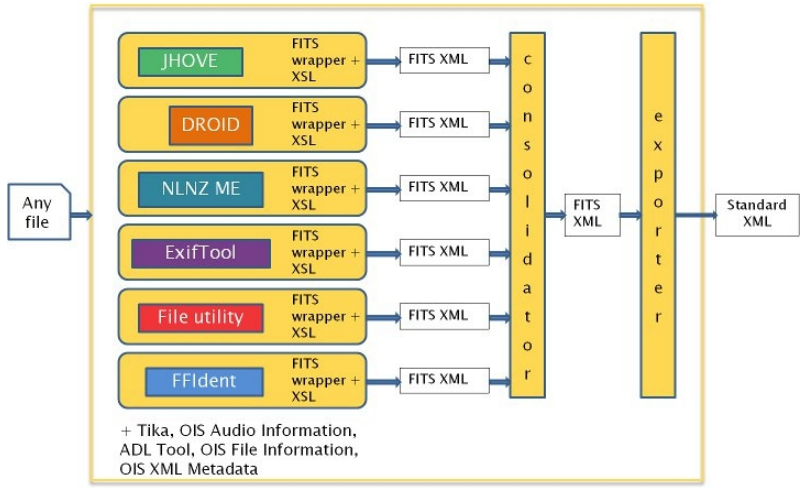
- mediainfo
  - Projet open-source, créé par Jérome Martinez
  - http://mediaarea.net/
  - Outil d'analyse des formats audio-video
    - Conteneur
      - MPEG-4, QuickTime, Matroska, AVI, MPEG-PS (y compris les DVD non protégés), MPEG-TS (y compris les Blu-ray non protégés), MXF, GXF, LXF, WMV, FLV, Real...
    - Tags
      - Id3v1, Id3v2, Vorbis comments, APE tags...
    - Vidéo
      - MPEG-1/2 Video, H.263, MPEG-4 Visual (DivX, XviD compris), H.264/AVC, Dirac...
    - Audio
      - MPEG Audio (MP3 compris), AC3, DTS, AAC, Dolby E, AES3, FLAC...
    - Sous-titres
      - CEA-608, CEA-708, DTVCC, SCTE-20, SCTE-128, ATSC/53, CDP, DVB Subtitle, Teletext, SRT, SSA, ASS, SAMI…



Les outils

- Tika
  - Porté par Apache software foundation
  - http://tika.apache.org
  - Utilise la technique des parsers
  - Utilise le mime/type pour l'identification
  - Permet la détection et l'extraction de métadonnées de différents formats de fichier (texte, image, audio, vidéo)
    - Permet l'extraction de métadonnées des formats bureautique ODF et OOXML
- File Information Tool Set (FITS)
  - Projet porté par l'université d'Harvard
  - http://projects.iq.harvard.edu/fits
  - Fédérateur de différents outils
  - Propose un format XML de sortie unifié

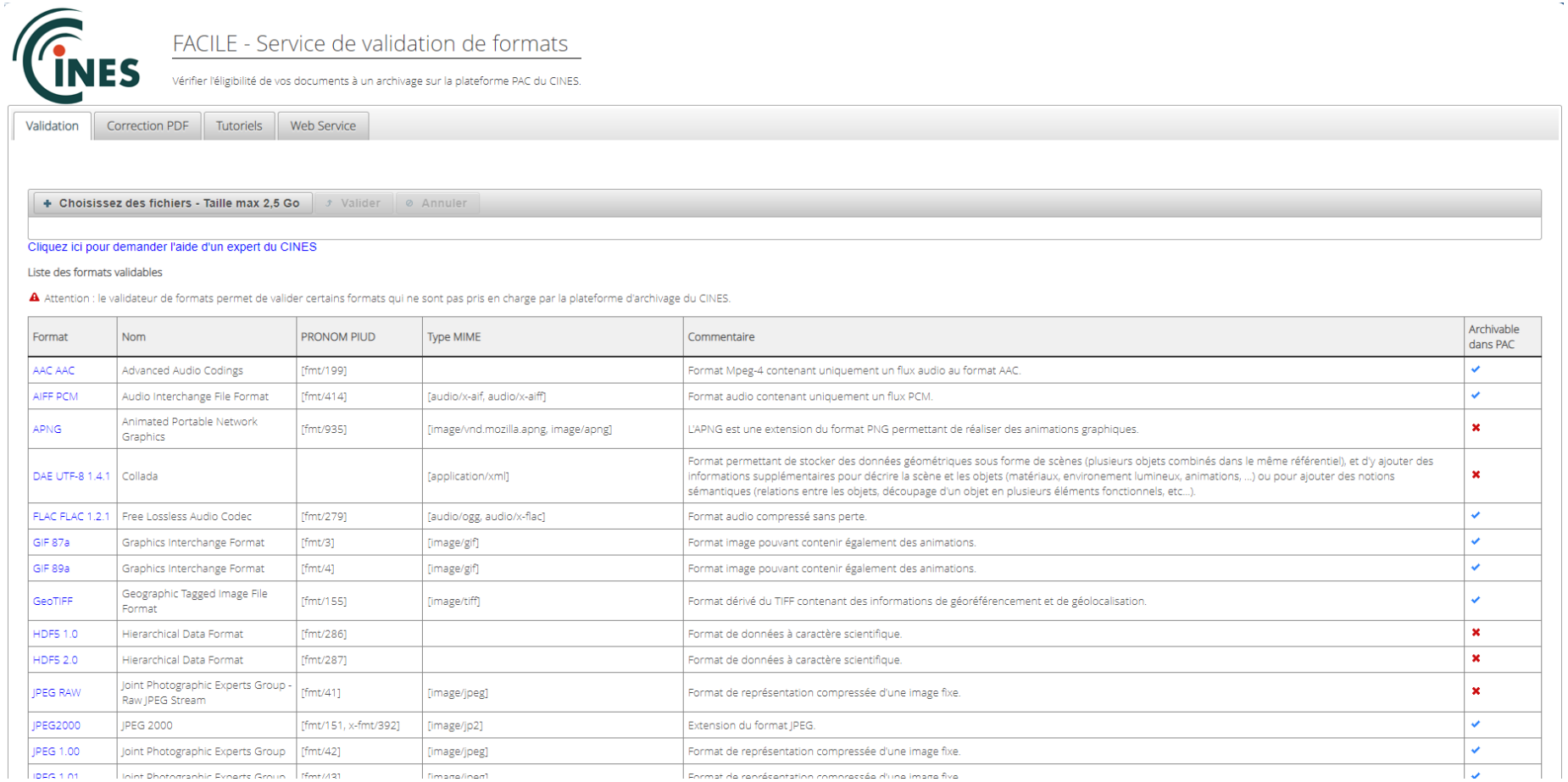




L’exemple FACILE au CINES

Format d'Archivage du CInes par anaLyse et Expertise

https://facile.cines.fr/



FACILE - Service de validation de formats

Vérifier l'éligibilité de vos documents à un archivage sur la plateforme PAC du CINES.

Validation | Correction PDF | Tutoriels | Web Service

+ Choisissez des fichiers - Taille max 2,5 Go | Valider | Annuler

Cliquez ici pour demander l'aide d'un expert du CINES

Liste des formats validables

Attention : le validateur de formats permet de valider certains formats qui ne sont pas pris en charge par la plateforme d'archivage du CINES.

Format	Nom	PRONOM PIUD	Type MIME	Commentaire	Archivable dans PAC
AAC AAC	Advanced Audio Codings	[fmt/199]		Format Mpeg-4 contenant uniquement un flux audio au format AAC.	✔
AIFF PCM	Audio Interchange File Format	[fmt/414]	[audio/x-aif, audio/x-aiff]	Format audio contenant uniquement un flux PCM.	✔
APNG	Animated Portable Network Graphics	[fmt/935]	[image/vnd.mozilla.apng, image/apng]	L'APNG est une extension du format PNG permettant de réaliser des animations graphiques.	✖
DAE UTF-8 1.4.1	Collada		[application/xml]	Format permettant de stocker des données géométriques sous forme de scènes (plusieurs objets combinés dans le même référentiel), et d'y ajouter des informations supplémentaires pour décrire la scène et les objets (matériaux, environement lumineux, animations, ...) ou pour ajouter des notions sémantiques (relations entre les objets, découpage d'un objet en plusieurs éléments fonctionnels, etc...).	✖
FLAC FLAC 1.2.1	Free Lossless Audio Codec	[fmt/279]	[audio/ogg, audio/x-flac]	Format audio compressé sans perte.	✔
GIF 87a	Graphics Interchange Format	[fmt/3]	[image/gif]	Format image pouvant contenir également des animations.	✔
GIF 89a	Graphics Interchange Format	[fmt/4]	[image/gif]	Format image pouvant contenir également des animations.	✔
GeoTIFF	Geographic Tagged Image File Format	[fmt/155]	[image/tiff]	Format dérivé du TIFF contenant des informations de géoréférencement et de géolocalisation.	✔
HDF5 1.0	Hierarchical Data Format	[fmt/286]		Format de données à caractère scientifique.	✖
HDF5 2.0	Hierarchical Data Format	[fmt/287]		Format de données à caractère scientifique.	✖
JPEG RAW	Joint Photographic Experts Group - Raw JPEG Stream	[fmt/41]	[image/jpeg]	Format de représentation compressée d'une image fixe.	✖
JPEG2000	JPEG 2000	[fmt/151, x-fmt/392]	[image/jp2]	Extension du format JPEG.	✔
JPEG 1.00	Joint Photographic Experts Group	[fmt/42]	[image/jpeg]	Format de représentation compressée d'une image fixe.	✔
JPEG 1.01	Joint Photographic Experts Group	[fmt/43]	[image/jpeg]	Format de représentation compressée d'une image fixe.	✔

FACILE

- Liste les formats de fichiers éligibles à l'archivage au CINES
- Permet
  - L'identification d'un format
  - La validation (par rapport au format identifié)
  - La correction éventuelle d'erreurs (format PDF uniquement)
  - L'assistance des experts du CINES
  - La formation aux bonnes pratiques (via des tutoriels)

Discussion

Vos formats

- Données numériques :
  - CSV (17) / Excel (10)
  - BDD relationnelles (5)
  - netCDF, XML (4)
  - geoTIFF, HDF, JSON ... (2)
  - ames, miniseed, FITS, et autres formats divers et variés ... (1)
- Données multimédias :
  - JPG, TIFF (4)
  - wav, MP3, MP4 (4)