

Qualité et préparation des données pour l'analyse – guide pratique

Table des matières

1. Introduction	3
2. Acquisition des données	3
2.1. Données externes	3
2.2. Données propres.....	4
2.3. Référence temporelle absolue.....	4
2.4. Chrono-timbre	5
2.5. Fréquence d'échantillonnage f et période d'échantillonnage T	5
2.5.1. Durée d'acquisition T	10
2.5.2. Changement d'heure.....	10
3. Préparation des données	10
3.1. Règles générales	11
3.2. Ajustement du fichier original	12
3.3. Format du chrono-timbre	12
3.4. Format des données.....	13
3.5. Nom des fichiers	14
3.6. Conclusion	17
A. Annexe	18
A. 1 Table des codes ASCII	18
A. 2 Outil pour la représentation graphique des mesures	19
A. 3 Glossaire.....	20

1. Introduction

Le projet AGID¹ a donné lieu à des collectes de données et à des campagnes de mesures. L'objectif était d'analyser divers processus de production et flux de matières et d'énergie, pris comme référence. Il s'agissait principalement de mesures relevées périodiquement, soit manuellement soit automatiquement, qui avaient été archivées dans un format quelconque, sans organisation particulière et sous une dénomination arbitraire. Dans de rares cas, les relevés étaient suffisamment complets et la qualité des données et leur organisation suffisantes pour permettre une utilisation ultérieure sans nécessiter de préparation excessive. Dans d'autres cas, plus nombreux, cette préparation fut plus coûteuse en temps que l'analyse à proprement parler. C'est la raison pour laquelle le présent guide a été élaboré. Il propose une procédure destinée à faciliter l'échange et la réutilisation des données et à améliorer leur qualité.

En l'absence de norme générale de formatage et d'organisation des données, nous proposons ici un format qui peut être lu par toutes les machines de traitement de texte, indépendamment du système d'exploitation et du logiciel utilisés. Une première organisation, sommaire, est obtenue grâce à une dénomination logique. Une organisation plus fine est possible ultérieurement, moyennant une importation dans des bases de données personnalisées.

2. Acquisition des données

L'acquisition des données s'entend comme le recueil systématique d'informations et leur enregistrement selon un modèle spécifique. Dans le cadre du projet AGID et des processus de production, des flux de matière et des flux d'énergie étudiés, les données sont variables dans le temps et comportent des informations sur l'état de divers sous-systèmes.

2.1. Données externes

Les données externes sont des données qui échappent à toute possibilité d'action. Il peut s'agir de données déjà archivées dans des fichiers ou des bases de données, ou bien de données récupérées directement à partir d'un système d'acquisition de mesures. Ce sont souvent les seules informations disponibles quant à l'état d'un processus. Dans de très rares cas, elles ont été enregistrées et archivées en continu, mais cette option est largement sous-exploitée. Les paramètres qui n'interviennent pas dans la régulation d'un système ne présentent évidemment pas d'intérêt pour les exploitants d'une installation. Selon ces derniers, l'entretien, la maintenance ou le paramétrage des équipements de mesure n'apportent pas de réelle plus-value pour l'exploitation, raison pour laquelle ils sont négligés. Ceci est souvent dû au fait que les installations fonctionnent avec des réglages de grandeurs prédéfinis, déterminés par approches successives à partir de l'expérience ; la spécificité de ces grandeurs est en effet telle que, même adaptée au mieux, aucune régulation ne parvient à un résultat vraiment satisfaisant (on citera pour exemple le cas des stations de relevage des réseaux d'assainissement pour lesquelles la

¹ AGID : Analyse et Gestion Intégrées et Durables des flux de matières et d'énergie en entreprise

composition et le débit d'aspiration ne sont pas maîtrisables). La plupart du temps, les données sont simplement conservées comme élément de preuve en cas d'incident, ou pour fournir des flux d'énergie ou de matières. C'est-à-dire qu'en temps normal, elles ne font l'objet d'aucun suivi. Lorsqu'elles sont fournies, elles le sont dans un format généralement imposé, qui ne peut être modifié. Dans tous les cas, un contrôle de vraisemblance s'impose : période de relevé, système d'acquisition de données (identification univoque) et relation à une référence temporelle absolue. L'exemple IV montre la préparation effectuée sur les relevés continus (relevés à distance) d'un compteur d'énergie d'un gestionnaire de réseau de distribution.

2.2. Données propres

Par données propres, on entend toute information qui se rapporte aux processus et que l'on produit soi-même. Il peut s'agir de valeurs lues sur l'écran d'un appareil de mesure et notées manuellement ou bien de valeurs enregistrées automatiquement par un enregistreur qui peut être paramétré et adapté en fonction des besoins propres.

2.3. Référence temporelle absolue

Même si cela n'est pas toujours nécessaire (en effet, les processus peuvent être chronologiquement indépendants les uns des autres), il est recommandé de rapporter les données recueillies à une référence temporelle absolue. Le temps est une grandeur mesurable, perçue par la conscience humaine comme une succession d'évènements ordonnés, selon une progression apparemment continue.²

C'est pourquoi, lorsqu'il s'agit de procéder à une analyse de fonctionnement ou lorsqu'on veut établir des relations entre différents jeux de données, l'utilisation d'une signature temporelle pour caractériser les données constitue une option importante (voir également **Error! Reference source not found.** et 3.3).

Une référence au temps absolu peut par exemple être obtenue au moyen d'une horloge GPS³ ou d'un récepteur DCF-77⁴ capable de recevoir le signal envoyé par le PTB Braunschweig⁵. Si l'on a affaire à un réseau d'instruments de mesure, on peut utiliser un serveur NTP⁶ qui constitue alors la base de temps pour l'ensemble des appareils. Un serveur NTP externe permet une liaison internet (par exemple serveur de temps du PTB Braunschweig : ptbtime1.ptb.de ou ptbtime2.ptb.de).

² d'après Wikipedia 2008

³ synchronisation de l'horloge de l'acquisition des données au moyen d'un récepteur GPS (Global Positioning System)

⁴ récepteur grandes ondes pouvant recevoir le signal radio de l'horloge atomique du PTB Braunschweig.

⁵ Physikalisch Technische Bundesanstalt Braunschweig

⁶ Serveur Network Time Protocol. Ce protocole a été développé afin de pouvoir synchroniser des ordinateurs à l'intérieur de réseaux locaux et de réseaux distants. Il est basé sur le protocole IP utilisé pour internet et est disponible pour tous les systèmes d'exploitation courants (source : PTB Braunschweig)

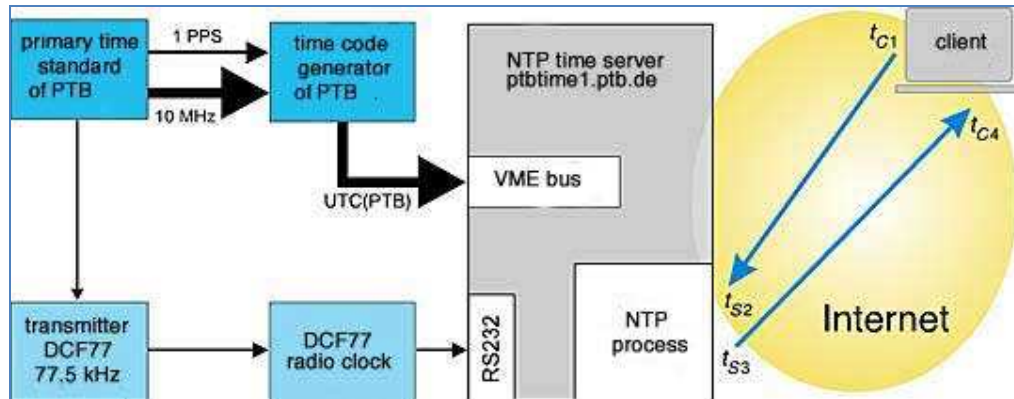


Figure 1 : Schéma fonctionnel d'une liaison internet serveur NTP-client, depuis le site du PTB Braunschweig (source : PTB Braunschweig)

2.4. Chrono-timbre

Chaque signal de mesure ou autre information doit pouvoir être horodaté. Ceci est réalisé au moyen de ce que l'on appelle une signature temporelle ou un chrono-timbre. La signature temporelle permet d'identifier des relations extérieures, soit à partir de systèmes d'acquisition de données différents soit à partir de sources de données distantes. Ces données peuvent correspondre à des grandeurs d'influence (des événements météorologiques pour la climatisation d'un bâtiment, par exemple), des échanges de flux de matière et d'énergie avec des systèmes voisins, qui pourraient agir sur le système considéré, ou encore des coûts et des prix de matières et d'énergies pouvant constituer une alternative à celles utilisées jusqu'alors, etc. La signature temporelle permet une synchronisation des données ; en d'autres termes, elle rend possible l'établissement d'une relation temporelle entre celles-ci.

2.5. Fréquence d'échantillonnage f et période d'échantillonnage T

Le nombre de mesures périodiques par intervalle de temps est appelé **fréquence d'échantillonnage** ou cadence d'échantillonnage (symbole : f , unité : Hertz [Hz]). L'intervalle de temps séparant deux mesures est appelé **période d'échantillonnage** (symbole : T^7 , unité : seconde [s]). Lorsque la période d'échantillonnage est inférieure à la seconde, c'est généralement la fréquence d'échantillonnage⁸ qui est indiquée. On fait la distinction entre la fréquence d'échantillonnage du signal proprement dit à mesurer et la fréquence d'échantillonnage correspondant à l'enregistrement des mesures du signal. Si on procède à des mesures limitées dans le temps, ces deux fréquences peuvent être identiques, chaque mesure étant aussi archivée. Dans le cas de mesures sur une durée plus longue, la fréquence d'échantillonnage du signal à mesurer est généralement élevée, les mesures faisant l'objet d'un enregistrement intermédiaire. A partir de ces

⁷ inverse de la fréquence d'échantillonnage : $T=1/f$, $f=1/T$

⁸ Exemple : une fréquence d'échantillonnage de 10 [Hz] correspond à une période $T= 1/ (10 [Hz]) = 0.1$ [s] = 100 [ms]

enregistrements provisoires, on calcule une valeur (moyenne arithmétique, moyenne glissante, intégrale, différentielle etc.) qui sera seule conservée.

Fréquence d'échantillonnage minimale f_{\min}

Souvent, le démarrage de la campagne de mesures est postérieur à la mise en service de l'appareil de mesure ; ou bien on s'intéresse à des données recueillies précédemment. Dans ces deux cas, l'optimisation de la fréquence d'échantillonnage n'est pas possible. Lorsque cette optimisation est possible (conception ou mise en œuvre d'un appareil ou d'un système), alors il convient de respecter le théorème de Nyquist-Shannon⁹ :

$$f_{\min} \geq 2 * f_{\text{signal}}$$

Ce théorème énonce que la fréquence d'échantillonnage doit être au moins égale à 2 fois la fréquence du signal mesuré. De cette manière, on est assuré d'une prise en compte de l'ensemble des variations dynamiques du système étudié et de l'absence d'altération du signal de mesure (p.ex. repliement du spectre, atténuation).

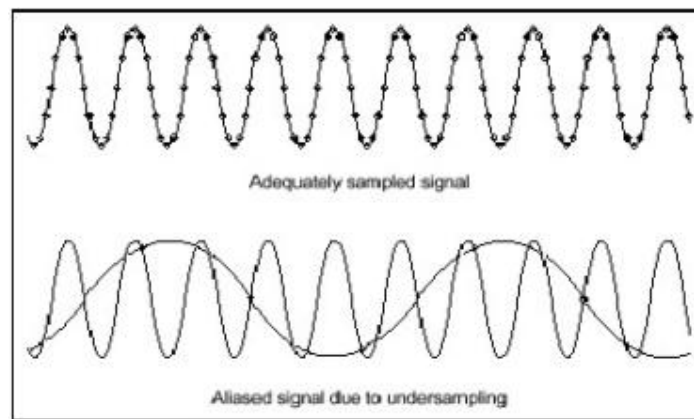


Figure 2 : Courbe du haut : fréquence d'échantillonnage correcte ; courbe du bas : fréquence d'échantillonnage trop faible (sous-échantillonnage ; anglais : undersampling). On a tracé la courbe reconstituée à partir des points de mesure et correspondant au repli du spectre d'une oscillation harmonique sinusoïdale. Elle conduit à une fréquence du signal 5 fois plus faible que la fréquence réelle. (Source : National Instruments).

Fréquence d'échantillonnage maximale f_{\max}

La fréquence d'échantillonnage maximale d'un appareil de mesure indique à quelle fréquence les mesures peuvent être délivrées. La fréquence d'échantillonnage du signal de sortie est ainsi déterminante. Dans le cas de systèmes de mesure intégrateurs comme les compteurs d'énergie, la fréquence d'échantillonnage des mesures est relativement élevée (de l'ordre du kilohertz pour les compteurs numériques, quasi-infinie¹⁰ pour les compteurs analogiques). La fréquence d'échantillonnage du signal de sortie, quant à elle, est beaucoup plus faible. Si nécessaire, la période d'échantillonnage¹¹ pour l'enregistrement

⁹ Harry Nyquist, physicien et Claude Elwood Shannon, mathématicien

¹⁰ Les facteurs limitants sont la masse des capteurs et le frottement des paliers mécaniques. Des modifications minimales influent sur le système de mesure sans pouvoir être restituées.

¹¹ NdT: le mélange permanent des fréquences et des périodes rend la lecture pénible. Il serait mieux d'être plus cohérent

des données par l'enregistreur peut à son tour être adaptée en fonction de la fréquence d'échantillonnage du signal de sortie. Elle est généralement relativement importante (un quart d'heure à une heure). La résolution¹² S de l'écran ou de l'interface de sortie de l'appareil de mesure est donnée par le nombre n d'impulsions ou de tours du disque par unité de comptage. La période d'échantillonnage minimale pour l'enregistrement¹³ peut être déterminée comme l'inverse du produit de la résolution par la puissance de raccordement maximale. Si l'on doit assurer l'enregistrement de la totalité des événements échantillonnés par l'instrument, l'enregistreur doit être réglé sur la fréquence d'échantillonnage minimale (période d'échantillonnage maximale).

I. Exemple d'ajustement de la fréquence d'échantillonnage minimale du signal d'entrée et détermination de la période d'échantillonnage minimale pour l'enregistrement

Cet exemple nous permet de montrer, à partir des caractéristiques d'un compteur électrique, comment la fréquence d'échantillonnage du signal d'entrée et la dynamique du signal à mesurer doivent être ajustés l'un par rapport à l'autre. Nous montrons ensuite quelle doit être la période d'échantillonnage minimale pour le stockage des données de manière à pouvoir rendre compte de la totalité des informations relatives à la variation des valeurs mesurées. L'acquisition des données est organisée de la manière suivante : un compteur d'énergie avec interface à impulsions est installé dans un circuit électrique entre une prise 230 V CA et le poste de consommation. Un compteur d'impulsions est raccordé à la sortie impulsionnelle du compteur d'énergie. Il fait la somme des impulsions délivrées par ce dernier et enregistre la valeur correspondante. Il est raccordé à son tour à un enregistreur qui l'interroge périodiquement pour connaître cette valeur, qu'il transforme en une valeur d'énergie, cette dernière étant ensuite archivée avec sa signature temporelle.

Caractéristiques techniques du compteur d'énergie :

Puissance de raccordement maximale : $P_{Max} = 24$ [kW]

Résolution de l'interface à impulsions : $S = 1\ 000$ impulsions / [kWh] =
1 impulsion / [Wh]

Durée minimale des impulsions : $T_{Impulse} = 30$ [ms] = 0.03 [s]

Il faut d'abord vérifier si le compteur d'impulsions est bien en mesure de détecter la totalité des impulsions. Pour cela, on calcule la fréquence d'échantillonnage minimale du compteur.

Fréquence des impulsions (à partir de leur durée minimale) :

$$f_{Impulse} = 1 / T_{Impulse} = 1 / (0.03[s]) = 33.3333 \text{ [Hz]}$$

Fréquence d'échantillonnage minimale f_{scan} du compteur d'impulsions pour une détection de la totalité des impulsions (à la sortie impulsionnelle) :

$$f_{scan} \geq 2 * f_{Impulse} \geq 66.67 \text{ [Hz]}$$

¹² correspond à la plus petite modification d'une valeur mesurée qu'il est possible d'observer

¹³ enregistrement obtenu par exemple en combinant un compteur d'énergie, qui émet des impulsions, un compteur d'impulsions, qui compte les impulsions et les somme, et un enregistreur, qui enregistre ces impulsions en tant que valeurs d'énergie.

Période d'échantillonnage maximale correspondante :

$$T_{scan} = 1/f_{scan} = 1/(66.67 \text{ [Hz]}) = 0.015 \text{ [s]}$$

En d'autres termes, lors du choix d'un compteur d'impulsions, il faut veiller à ce que celui-ci autorise une fréquence d'échantillonnage minimale de 66.67 [Hz] pour que toutes les impulsions du compteur d'énergie puissent être prises en compte. Pour chaque impulsion détectée, une mémoire non volatile du compteur d'impulsions est incrémentée d'une unité. La valeur stockée dans la mémoire est conservée même en cas d'interruption de l'alimentation. L'énergie est obtenue en divisant cette valeur par la résolution S de l'interface impulsionnelle.

Si l'on veut exploiter au maximum la résolution de l'interface impulsionnelle pour la puissance raccordée maximale - c'est-à-dire si l'on veut que la moindre modification soit enregistrée -, alors la fréquence d'échantillonnage maximale (ou la période d'échantillonnage minimale) pour l'enregistrement se calcule comme suit :

Fréquence d'échantillonnage maximale par impulsion pour la puissance raccordée maximale :

$$f_{max} = S * P_{Max} / (\text{impulsions} * 3600 \text{ [s]})$$

$$f_{max} = 1000 \text{ impulsions} * 24 \text{ [kW]} / (\text{kWh} * \text{impulsions} * 3600 \text{ [s]/h}) = 6.67 \text{ [Hz]}$$

Période d'échantillonnage minimale : $T_{min} = 1/f_{Max}$

$$T_{min} = 1/(6.67 \text{ [Hz]}) = 0.15 \text{ [s]}$$

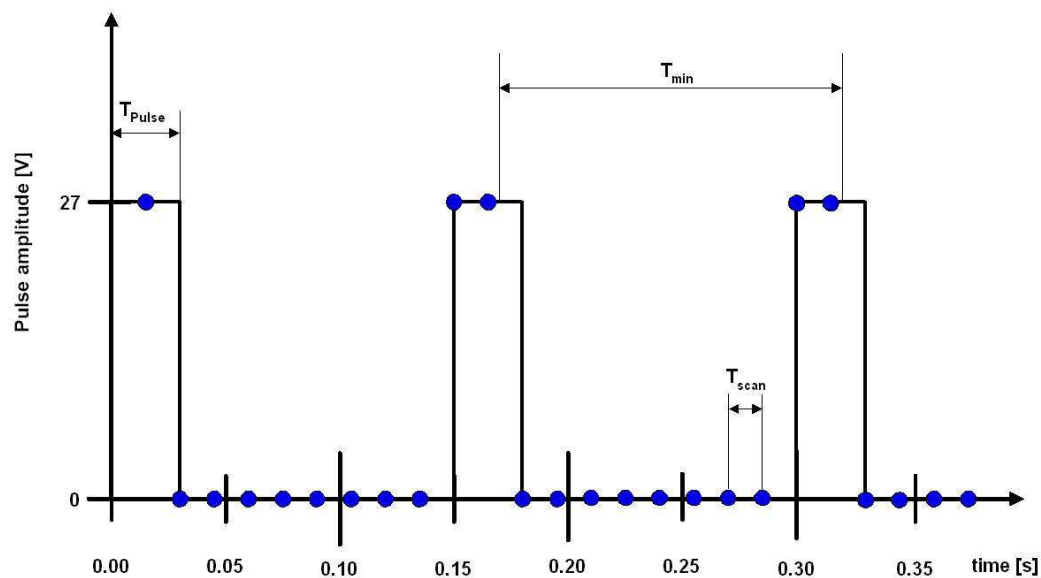


Figure 3 : Représentation des périodes d'échantillonnage calculées dans l'exemple proposé (compteur d'énergie avec interface S0, puissance raccordée maximale). La durée des impulsions est supposée constante $T_{pulse} = 30$ millisecondes. La période d'échantillonnage maximale du compteur d'impulsions (points bleus) raccordé à l'interface S0 vaut $T_{scan} = 15$ millisecondes. La période d'échantillonnage minimale de l'enregistreur raccordé à la sortie du compteur d'impulsions vaut $T_{min} = 150$ millisecondes (0,15 secondes).

Fréquence d'échantillonnage optimale

L'optimisation de la fréquence d'échantillonnage se traduit le plus souvent par un compromis. La fréquence maximale permet d'obtenir un maximum de détails quant au comportement dynamique d'un système. Mais il est clair qu'une fréquence d'échantillonnage élevée conduit rapidement aux limites des capacités de stockage même les plus importantes dès lors que l'on enregistre chacune des valeurs. C'est pourquoi il est important de connaître le comportement dynamique du système considéré et de se demander si les informations sont nécessaires et avec quelle résolution. Ceci permet d'optimiser l'espace mémoire disponible et facilite le traitement a posteriori des données. Dans l'exemple numérique I, il n'est pas intéressant de retenir la fréquence d'échantillonnage maximale pour l'enregistrement des données si, pour une application particulière, la puissance mesurée représente une fraction seulement de la puissance maximale ou si elle ne varie que lentement avec le temps. Choisir des périodes d'échantillonnage courtes peut parfaitement être judicieux, y compris dans le cas de campagnes de mesures de longue durée ou pour de la surveillance dès lors que, en plus de l'état du système, on veut étudier son comportement dynamique, c'est-à-dire les variations imprévisibles des signaux dans le temps. Par expérience, pour des durées d'acquisition étendues, la période d'échantillonnage pour l'enregistrement des mesures est comprise entre une minute et une heure.

En cas de nécessité, il est possible d'homogénéiser a posteriori les fréquences d'échantillonnage de différentes séries de mesures. C'est le cas lorsque l'analyse ou la modélisation imposent une même fréquence d'échantillonnage pour toutes les séries - appelées ci-après séries chronologiques. Il faut alors déterminer la plus grande fréquence d'échantillonnage commune à toutes ces séries. La réduction a posteriori de la fréquence d'échantillonnage appelle les considérations suivantes : dans le cas de grandeurs intégrales - énergie ou niveau de remplissage, par exemple - l'heure et la date du dernier point de mesure doivent figurer dans la signature temporelle. Dans le cas de grandeurs représentant des valeurs instantanées - débit, puissance, température, par exemple -, la réduction de la fréquence d'échantillonnage fait appel aux valeurs moyennes. Par conséquent, l'heure et la date correspondantes se situent à mi-chemin entre celles de la première et de la dernière mesure. Si maintenant on rapproche ces deux types de grandeurs, l'instant commun entre ceux-ci est celui à la fin de la période d'échantillonnage commune.

II. Exemple : synchronisation de séries chronologiques avec des périodes d'échantillonnage différentes

La série chronologique d'une mesure d'énergie a été enregistrée avec une période d'échantillonnage d'une minute. On veut la synchroniser avec une autre série, pour laquelle la période d'échantillonnage est de 1 heure. Pour cela, on fait la somme des valeurs obtenues minute après minute entre le début et la fin de chaque heure, on convertit la valeur obtenue (60 watt-minutes correspondant à 1 wattheure), et on lui attribue une nouvelle signature temporelle. Les deux séries ont maintenant des signatures temporelles qui concordent : les mesures sont synchronisées.

2.5.1. Durée d'acquisition T

La durée d'acquisition s'entend comme l'intervalle de temps pendant lequel des données ont été enregistrées en continu. Les manques dus par exemple à l'absence de données peuvent être comblés par extrapolation ou, comme dans le cas des données météorologiques, par importation des données existantes provenant de la station météorologique (par exemple de l'aéroport du Findel, Luxembourg). La continuité des valeurs présente des avantages en termes de traitement des données, d'analyse et de modélisation. Les petites imprécisions liées à leur suppression ne sont pas identifiables au niveau de la valeur annuelle cumulée.

2.5.2. Changement d'heure

Si la durée d'acquisition comporte un ou plusieurs passages de l'heure d'été à l'heure d'hiver ou inversement, les données manquantes ou les recouvrements correspondants doivent être éliminés. A notre longitude, on recommande d'utiliser comme base de temps l'UTC¹⁴ ou l'UTC + 1[h] - correspondant à l'heure normale d'Europe centrale. Pour faciliter le traitement des données, il convient dans tous les cas d'éviter un changement automatique de l'heure d'été à l'heure d'hiver - correction de la "perte" de données pendant une heure au passage de l'heure d'hiver à l'heure d'été ("saut" de 2h00 à 3h00) et correction du recouvrement des données au passage de l'heure d'été à l'heure d'hiver (recul de 3h00 à 2h00).

3. Préparation des données

Dans ce chapitre, nous présentons un format de données clairement structuré, simple à utiliser pour les traitements ultérieurs. Nous montrons en outre, à l'aide d'un exemple, comment traiter un format spécifique de données, non parfaitement optimisé, de manière à le transformer en un format générique facile à utiliser. Ceci se veut une piste dans le cas de l'établissement d'un nouveau système d'acquisition de données ou de la modification d'un système existant. Après plusieurs années d'activité dans le domaine de la métrologie et de la régulation, le CRP Henri Tudor confirme que les données externes sont régulièrement fournies dans des formats difficiles à utiliser de manière générale. Pour des utilisations isolées ou pour un système spécifique, ceci peut ne pas présenter d'inconvénient, mais lorsqu'il s'agit de rassembler des données transversales provenant de plusieurs systèmes, il est avantageux d'observer quelques règles générales. Dans l'état actuel de nos connaissances, il n'existe malheureusement pas encore de norme internationale qui réponde par ailleurs à l'ensemble des exigences du traitement de données. La préparation des données brutes - ou, mieux, "originales" - est plus ou moins difficile selon le cas. Une telle préparation est généralement nécessaire chaque fois qu'il faut synchroniser des données provenant de sources différentes, afin de pouvoir les analyser ensuite conjointement.

¹⁴ UTC = universal time coordinated (anglais), c'est-à-dire temps universel coordonné, précédemment désigné par GMT, Greenwich mean time.

3.1. Règles générales

Toutes les données sont enregistrées au format ASCII dans un fichier texte tabulé¹⁵, avec l'extension ".txt". Ceci a pour avantage que les données pourront être lues ultérieurement avec n'importe quel logiciel de traitement de texte et qu'elles ne sont pas liées à une application particulière qui pourrait un jour ne plus être disponible. Toutes les informations complémentaires concernant les mesures (que l'on appelle les métadonnées) sont écrites dans l'en-tête du jeu de données considéré. Ceci facilite l'identification des unités de mesure et des spécifications des appareils et constitue une solution simple, en cas d'importation dans d'autres applications informatiques, pour affecter de manière correcte des propriétés (surtout les unités correctes) aux données. Cet en-tête doit être séparé du contenu à proprement parler par un signe clair et sans équivoque, que l'on ne retrouvera normalement pas dans le corps des données. Il convient de ne pas utiliser de signe qui pourrait être utilisé comme variable d'environnement ni de signe ne pouvant être affiché dans un éditeur de texte courant, tel le retour à la ligne. Le mieux est d'utiliser un caractère et de le répéter afin d'éviter qu'il ne soit confondu avec un caractère qui pourrait avoir un autre sens en langage de programmation.

Le signe "dièse" #¹⁶ ou l'astérisque *, signes de commentaires usuels, utilisés sous la forme ### ou ****, ou un texte descriptif clair peuvent ainsi être utilisés pour séparer l'en-tête.

III. Exemple : en-tête type d'un fichier de données

Contenu de l'en-tête d'un fichier de mesures :

```
***Meteorologic Measurements: Data logger Keithley KE 2701E Serialnumber  
975668      Firmware Revision B09***
```

Time stamp information:

Data is recorded with a time stamp in CET (central European Time) which corresponds to GMT (Greenwich Mean Time) +1 hour.

the Wind direction in column 8 is calculated in moving average of the last 10 minutes. the first ten values are not related to the moving averages of the day before. Actually there is no need for this measure.

Description of sensor connections, Keithley channel connection and calculation:

```
1.) Date trigger CET (Central European Time) 2.) Time stamp 3.) 101INTCHAN:  
Direct Irradiance [W/m²] 4.) 102INTCHAN: Global Horizontal Irradiance  
[W/m²] 5.)103INTCHAN: Global Tilted Irradiance (30° tilt, 180° azimuth)  
[W/m²] 6.) 104INTCHAN: Diffuse Irradiance [W/m²] 7.) 105INTCHAN: Wind  
Speed [m/s] 8.) 106INTCHAN: Wind Direction [°]moving average of wind  
direction (10 minutes mean) 9.) 107INTCHAN: Relative Air Humidity [% rH]  
10.) 108INTCHAN: Ambient Temperature [° C] 11.) 109INTCHAN: Barometric  
Pressure [hPa] 12.) 110INTCHAN: Bodytemperature Pyrhelimeter [°C] 13.)  
111INTCHAN: Bodytemperature Pyranometer global horizontal [°C] 14.)  
112INTCHAN: Bodytemperature Pyranometer Global Tilted [°C] 15.) 113INTCHAN:  
Bodytemperature Pyranometer diffuse [°C] 16.) 114INTCHAN: Bodytemperature  
Rotronic Hygroclip [°C]
```

¹⁵ En informatique, on utilise souvent l'expression anglaise "tab separated"

¹⁶ désigné couramment en informatique par le terme anglais "hash"

```
#data table header#date time text Time stamp CH01 Dir Irr [W/m2] CH02  
Glob Hor Irr [W/m2] CH03 Glob 30°Tilt Irr [W/m2] CH04 Diff Irr [W/m2]  
CH05 Wind Speed [m/s] CH06 Wind Dir [°] CH07 Rel Air Hum [% rH] CH08  
Amb Temp [° C] CH09 Bar Press [hPa] CH10 temp Pyrhel[°C] CH11 temp  
Pyr glob hor[°C] CH12 temp Pyr Glob 30° [°C] CH13 temp Pyr diff [°C]  
CH14 temp Hygroclip [°C] Mean wind direction [°] glob hor irradiation  
[Wh/m2]
```

La marque de séparation entre l'en-tête et le corps du fichier est constituée ici par le passage surligné en jaune : `#data table header#`. Le texte qui suit représente les titres des colonnes de données.

3.2. Ajustement du fichier original

Avant toute manipulation des données originales, il convient de toujours effectuer une copie de sauvegarde. Toute modification des données originales devrait être documentée afin de pouvoir identifier toute erreur éventuelle et la corriger le cas échéant. Cette documentation peut servir par ailleurs à l'auteur des données originales pour mettre à disposition les données futures dans la forme souhaitée. La partie du fichier texte contenant les données à proprement parler ne doit pas comporter de signes répétés entre les données (en particulier pas de caractères non imprimables comme des tabulations, des retours à la ligne ou des sauts de lignes). Ils risqueraient d'entraîner une modification involontaire, voire inaperçue, de la position des données lors de leur importation dans un tableur. Le format ASCII ne comportant qu'un nombre limité de caractères typographiques, de nombreux signes ne peuvent pas être représentés. Voir le tableau en annexe.

3.3. Format du chrono-timbre

Quelles que soient les séries de données considérées, le format retenu pour la signature temporelle est le suivant (selon ISO 8601:2004 et EN 28601) :

`YYYY-MM-DD\shh:mm:ss`

`YYYY` : année, 4 chiffres, par exemple 2007

`MM` : mois, 2 chiffres, par exemple février = 02

`DD` : jour du mois, 2 chiffres, par exemple 09

`\s` : espace (anglais informatique : "backslash s"), code ASCII décimal 32

`hh` : heure, 2 chiffres, de 00 à 23

`mm` : minutes, 2 chiffres, de 00 à 59

`ss` : secondes, 2 chiffres, de 00 à 59

Ce format garantit un horodatage univoque des données lors de leur importation au format ASCII dans un logiciel d'analyse ou une base de données (LabVIEW, Matlab, MS Office etc.). Un autre format pour la signature temporelle pourrait être celui du temps Unix ; il s'agit d'un entier 32 bits (en secondes depuis le 1er janvier 1970 00:00 UTC). Mais ce format ne doit pas être retenu pour deux raisons : 1) rares sont ceux qui, en voyant un nombre entier, sont capables de distinguer sans problèmes qu'il s'agit d'une date et d'une heure ; 2) l'entier 32 bits atteindra sa valeur maximale le 19 janvier 2038 ;

au-delà, il y aura réécriture et l'on s'attend à un problème analogue à celui du bug de l'an 2000.

L'utilisation du format proposé ici pour la signature temporelle devrait éliminer les problèmes de reconnaissance dans les applications informatiques les plus usuelles. On recommande néanmoins de porter un regard critique sur le contenu de la signature. Ceci vaut en particulier pour le changement de jour et la dernière / la première indication temporelle de la journée. Dans certains jeux de données, la journée se termine à 00:00:00, dans d'autres à 24:00:00. Lors de l'importation dans un logiciel, il est fort probable que ces deux heures seront interprétées de manière erronée, avec erreur sur la signature temporelle. On notera que certaines compagnies de l'industrie gazières utilisent un découpage particulièrement curieux : la journée commence à 06:00:00 et finit à 05:59:59 le lendemain. Pour éviter toute erreur d'interprétation par le logiciel, on retiendra 00:00 pour le début de la journée et 23:59:59 pour la fin.

3.4. Format des données

Le séparateur décimal utilisé pour les données (mesures) à proprement parler est le point. On évitera si possible d'avoir 3 chiffres derrière la virgule afin d'éviter la confusion avec le point utilisé couramment comme séparateur des milliers en allemand par exemple (arrondir par exemple 123.456 à 123.46 ou ajouter un "0" : 123.4560). Avant d'importer des données, vérifier le réglage des options régionales de l'ordinateur. La séparation des milliers (par exemple 1'000 au lieu de 1000 en allemand) est très intéressante pour la lisibilité des grands nombres, mais elle n'est pas retenue ici pour éviter les erreurs d'interprétation dans le cas d'un échange de données avec des ordinateurs ne possédant pas les mêmes réglages des paramètres régionaux. Les mesures ainsi préparées sont écrites sans unité sur une même ligne à la suite de la signature temporelle. Elles en sont séparées par une tabulation [\t] (code ASCII décimal 09, désigné par TAB, symbole \t). A la fin de chaque ligne, le retour à la ligne est obtenu avec [\r\n]¹⁷ : retour chariot (code ASCII décimal 13, désignation CR, symbole \r) et saut de ligne (code ASCII décimal 10, désignation LF, symbole \n).

Exemple de série de mesures (les tabulations, espaces et retours à la ligne ne sont pas visibles) :

2007-11-14 16:00:00	1352.67
2007-11-14 16:15:00	1374.03
2007-11-14 16:30:00	1330.45

Avec ce format, dans la plupart des applications informatiques du type tableur ou traitement de texte, les données sont représentées sous forme d'un tableau à 2 entrées et peuvent ainsi être facilement utilisées. On notera par ailleurs que le nombre de lignes des tableurs est souvent limité à $2^{16} = 65'536$ et le nombre de colonnes à $2^8 = 256$. Si on doit utiliser des jeux de données contenant un nombre plus important de valeurs, on recommande de les importer dans une base de données.

¹⁷ \r\n : carriage return line feed, en français : retour chariot – saut de ligne ; dans le format texte de Windows, constitue le signe du retour à la ligne.

3.5. Nom des fichiers

Il est toujours avantageux de disposer de fichiers de mesures dont le nom permet d'identifier le type de mesures, le moment auquel elles ont été effectuées et l'opérateur. Afin de permettre un classement chronologique¹⁸ à l'intérieur du dossier dans lequel ces fichiers sont stockés, nous suggérons la forme suivante :

date_heure_point de mesure_opérateur.XXX

Date : YYYYMMDD

Heure : hhmss (peut être omis, selon le cas)

Point de mesure : abréviation ou acronyme du point de mesure ou du capteur, par exemple "tempchaudi" pour "température de la chaudière" ou "meteoFind" pour "station météorologique du Findel"

Opérateur : personne, entreprise ou organisme en charge du relevé.

.XXX : extension du fichier¹⁹ Elle est séparée du reste du nom du fichier par un point (exemple : .txt ou .xls).

IV. Exemple de préparation des données

Nous montrons ici, pas à pas, à l'aide d'un exemple concret, comment les données originales ont été transposées dans un format utilisable ultérieurement. Les données ont été fournies par la ville d'Esch sur Alzette. A priori, la présentation est claire et permet de travailler avec un tableur. Mais pour réaliser une synchronisation avec d'autres données, relevées en parallèle en d'autres points du même système, ce format ne peut pas être utilisé dans cet état. Le format de sortie est présenté dans la Figure 4. Il s'agit d'un tableau à 2 entrées. Les colonnes correspondent aux heures, les lignes aux dates. La présentation est interrompue au 28.10.2007, date du passage de l'heure d'été à l'heure d'hiver. La dernière heure de la journée est indiquée par 00:00. Il s'agit, à partir de cet exemple, de présenter une méthode permettant la transformation de ce tableau à 2 entrées, dont la présentation ne se prête pas de manière optimale à un traitement ultérieur, en une série chronologique avec une signature temporelle correcte pour chaque ligne et la mesure correspondante.

1. Chaque mois, les données sont envoyées par e-mail par la ville d'Esch/Alzette. Le nom des fichiers est identique tous les mois. Les fichiers doivent par conséquent être renommés afin d'éviter leur écrasement, mais surtout pour faciliter l'identification ultérieure d'un grand ensemble de données. Les noms d'origine des fichiers sont conservés dans une archive d'e-mail. Dans le présent exemple, le fichier "Generator1.csv" est renommé "200710_BHKW_1.csv".

¹⁸ Le type de classement optimal (alphabétique ou chronologique) a déjà fait l'objet de nombreuses discussions. Etant donné que les mesures s'étendent généralement sur une certaine durée, ce type de classement s'est avéré le plus commode. Les noms de fichier basés sur un classement alphabétique peuvent également être triés, pour un même nom, par date et heure, si celles-ci sont ajoutées à la suite du nom.

¹⁹ anglais informatique : filename extension

2. On vérifie d'abord si, dans le jeu de données mensuelles considéré, on est passé de l'heure d'hiver (heure normale d'Europe centrale - HNEC) à l'heure d'été (heure avancée d'Europe centrale - HAEC) (voir Figure 4). Ceci figure dans une ligne spéciale du jeu de données. Pour pouvoir être comparées à d'autres séries temporelles, mais surtout pour éviter les manques et les recouvrements lors de la modification du format de date, ces données doivent être converties au même format HNEC.

	00:15 Uhr	00:30 Uhr	00:45 Uhr	01:00 Uhr	01:15 Uhr	01:30 Uhr	01:45 Uhr	02:00 Uhr	02:15 Uhr	02:30 Uhr	22:15 Uhr	22:30 Uhr	22:45 Uhr	23:00 Uhr	23:15 Uhr	23:30 Uhr	23:45 Uhr	00:00 Uhr					
01.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
02.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
03.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
04.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
05.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
06.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
07.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
08.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
09.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
10.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
11.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
12.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
13.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
14.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
15.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
16.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
17.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
18.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
19.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
20.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
21.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
22.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
23.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	246,00	0	0	0	0	0	0	0					
24.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
25.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
26.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
27.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
(Zehnmstellung)	00:15 Uhr	00:30 Uhr	00:45 Uhr	01:00 Uhr	01:15 Uhr	01:30 Uhr	01:45 Uhr	02:00 Uhr	02:15 Uhr	02:30 Uhr	21:15 Uhr	21:30 Uhr	21:45 Uhr	22:00 Uhr	22:15 Uhr	22:30 Uhr	22:45 Uhr	23:00 Uhr	23:15 Uhr	23:30 Uhr	23:45 Uhr	00:00 Uhr	
28.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	00:15 Uhr	00:30 Uhr	00:45 Uhr	01:00 Uhr	01:15 Uhr	01:30 Uhr	01:45 Uhr	02:00 Uhr	02:15 Uhr	02:30 Uhr	22:15 Uhr	22:30 Uhr	22:45 Uhr	23:00 Uhr	23:15 Uhr	23:30 Uhr	23:45 Uhr	00:00 Uhr					
29.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
30.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0					
31.10.07 00:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	238,65	0	0	0	0

Figure 4 : Présentation des données brutes des courbes de charge de l'unité de cogénération. Puissance du générateur, octobre 2007, avec passage de l'heure d'été à l'heure d'hiver

3. Une étape intermédiaire consiste à exporter les fichiers CSV convertis en HNEC (fichiers textes tabulés).
4. A l'aide d'un algorithme approprié, les fichiers textes sont mis dans un nouveau format, renommés et enregistrés à nouveau sous forme de fichiers textes tabulés. Les mesures provenant du tableau original à 2 entrées sont ajoutées les unes après les autres, avec leur signature temporelle. Compte tenu que, pour les données brutes, on a attribué à la dernière valeur de la journée l'heure 00:00:00, et que la première mesure, le lendemain, est effectuée à 00:15:00, la dernière valeur de la journée (avec l'heure 00:00:00) doit également être attribuée à la date du jour suivant. Dans le cas contraire, la série chronologique présente une discontinuité. L'algorithme utilisé pour le traitement des données brutes a été développé sous LabVIEW²⁰, parce que cela permet, sans trop de difficultés, d'automatiser le traitement (voir Figure 5). Un traitement manuel ne saurait être recommandé que dans le cas où le nombre de données est limité et où le traitement n'aura pas à être renouvelé. Le traitement manuel suppose d'être particulièrement attentif au risque, sinon, de voir des erreurs se glisser dans les données. C'est là déjà une bonne raison pour préférer un traitement informatisé. Les fichiers ainsi obtenus lors de cette étape intermédiaire sont appelés fichiers de travail ou fichiers temporaires. Ils sont dénommés en conséquence et archivés pour une utilisation ultérieure.

²⁰ environnement de programmation graphique développé par National Instruments, USA, spécialement adapté pour l'automatisation, pour l'acquisition des données ainsi que pour le traitement des données.

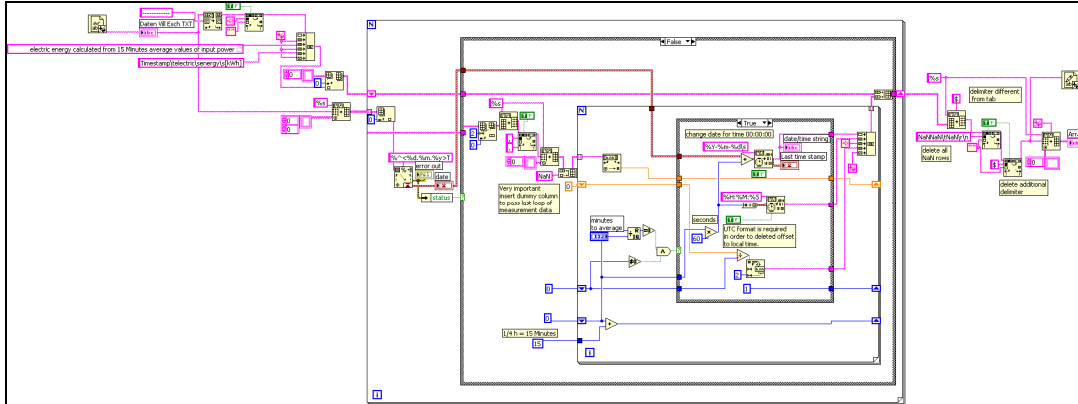


Figure 5 : Algorithme pour le traitement des données brutes de la ville d'Esch/Alzette (gestionnaire des réseaux de distribution d'énergie) dans le langage de programmation graphique LabVIEW

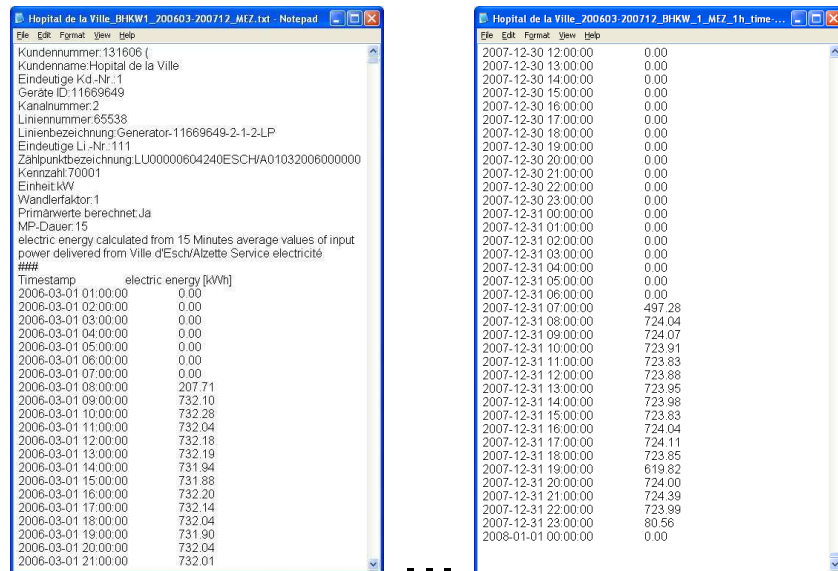


Figure 6 : Fichier chronologique après traitement des données : courbe de charge de l'unité de cogénération. Puissance du générateur, octobre 2007, avec signature temporelle uniforme (HNEC)

5. Le contenu des différents fichiers de travail est copié dans un fichier global et enregistré pour la suite de l'analyse sous forme d'un fichier chronologique final (voir Figure 6). Ce fichier permet de présenter l'ensemble des mesures dans les colonnes d'un tableau, de les comparer et de les analyser.

3.6. Conclusion

Les différents sous-projets du projet AGID ont montré que la préparation des mesures et des données de consommation existantes pouvait prendre un temps excessif. D'autres projets ont conduit à des constatations analogues. C'est pourquoi, lors de l'étude d'un nouveau système d'acquisition de données ou lors de la définition d'une campagne de mesures, il convient de veiller systématiquement à réduire le temps de préparation des données, par une configuration adaptée du système. Là où cela n'est pas possible – dans le cas de systèmes existants –, les procédures et les exemples présentés dans ce guide constituent des repères pour un traitement adéquat. Il est prévu d'actualiser régulièrement le présent document, à mesure du retour d'expérience des projets à venir.

A. Annexe

A. 1 Table des codes ASCII

Dez	Hex	Okt	Zeichen	Dez	Hex	Okt	Zeichen	Dez	Hex	Okt	Zeichen	Dez	Hex	Okt	Zeichen
0	0x00	0	NUL	32	0x20	40	SP	64	0x40	100	@	96	0x60	140	`
1	0x01	1	SOH	33	0x21	41	!	65	0x41	101	A	97	0x61	141	a
2	0x02	2	STX	34	0x22	42	"	66	0x42	102	B	98	0x62	142	b
3	0x03	3	ETX	35	0x23	43	#	67	0x43	103	C	99	0x63	143	c
4	0x04	4	EOT	36	0x24	44	\$	68	0x44	104	D	100	0x64	144	d
5	0x05	5	ENQ	37	0x25	45	%	69	0x45	105	E	101	0x65	145	e
6	0x06	6	ACK	38	0x26	46	&	70	0x46	106	F	102	0x66	146	f
7	0x07	7	BEL	39	0x27	47	'	71	0x47	107	G	103	0x67	147	g
8	0x08	10	BS	40	0x28	50	(72	0x48	110	H	104	0x68	150	h
9	0x09	11	TAB	41	0x29	51)	73	0x49	111	I	105	0x69	151	i
10	0x0A	12	LF	42	0x2A	52	*	74	0x4A	112	J	106	0x6A	152	j
11	0x0B	13	VT	43	0x2B	53	+	75	0x4B	113	K	107	0x6B	153	k
12	0x0C	14	FF	44	0x2C	54	,	76	0x4C	114	L	108	0x6C	154	l
13	0x0D	15	CR	45	0x2D	55	-	77	0x4D	115	M	109	0x6D	155	m
14	0x0E	16	SO	46	0x2E	56	.	78	0x4E	116	N	110	0x6E	156	n
15	0x0F	17	SI	47	0x2F	57	/	79	0x4F	117	O	111	0x6F	157	o
16	0x10	20	DLE	48	0x30	60	0	80	0x50	120	P	112	0x70	160	p
17	0x11	21	DC1	49	0x31	61	1	81	0x51	121	Q	113	0x71	161	q
18	0x12	22	DC2	50	0x32	62	2	82	0x52	122	R	114	0x72	162	r
19	0x13	23	DC3	51	0x33	63	3	83	0x53	123	S	115	0x73	163	s
20	0x14	24	DC4	52	0x34	64	4	84	0x54	124	T	116	0x74	164	t
21	0x15	25	NAK	53	0x35	65	5	85	0x55	125	U	117	0x75	165	u
22	0x16	26	SYN	54	0x36	66	6	86	0x56	126	V	118	0x76	166	v
23	0x17	27	ETB	55	0x37	67	7	87	0x57	127	W	119	0x77	167	w
24	0x18	30	CAN	56	0x38	70	8	88	0x58	130	X	120	0x78	170	x
25	0x19	31	EM	57	0x39	71	9	89	0x59	131	Y	121	0x79	171	y
26	0x1A	32	SUB	58	0x3A	72	:	90	0x5A	132	Z	122	0x7A	172	z
27	0x1B	33	ESC	59	0x3B	73	;	91	0x5B	133	[123	0x7B	173	{
28	0x1C	34	FS	60	0x3C	74	<	92	0x5C	134	\	124	0x7C	174	
29	0x1D	35	GS	61	0x3D	75	=	93	0x5D	135]	125	0x7D	175	}
30	0x1E	36	RS	62	0x3E	76	>	94	0x5E	136	^	126	0x7E	176	~
31	0x1F	37	US	63	0x3F	77	?	95	0x5F	137	_	127	0x7F	177	DEL

Figure 7 : Table des codes ASCII. Les caractères 0 -31 et 127 (jaunes) sont appelés caractères de contrôle ; ils ne sont pas imprimables.

A. 2 Outil pour la représentation graphique des mesures

Un programme a été écrit sous LabVIEW pour permettre une première analyse visuelle des données du projet AGID ; il permet une représentation graphique de l'ensemble des mesures de séries chronologiques avec le format décrit au paragraphe 3. Cette représentation graphique offre des fonctions de zoom, de masque, de translation des axes etc. pour l'analyse (voir Figure 8 et Figure 9). Un exécutable de ce programme peut, si besoin, être demandé auprès du CRTE.

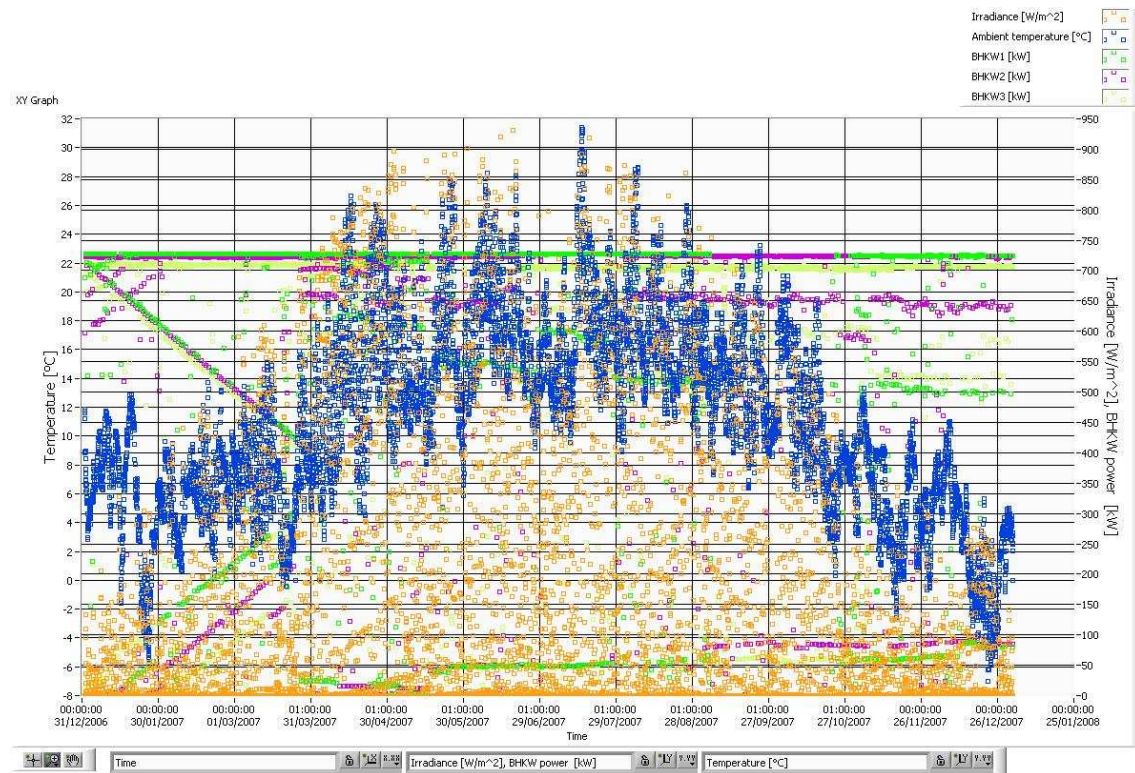


Figure 8 : Récapitulatif de l'ensemble des mesures disponibles sur l'année, relevées en continu, avec une période d'échantillonnage de 1 heure (1 [h]).

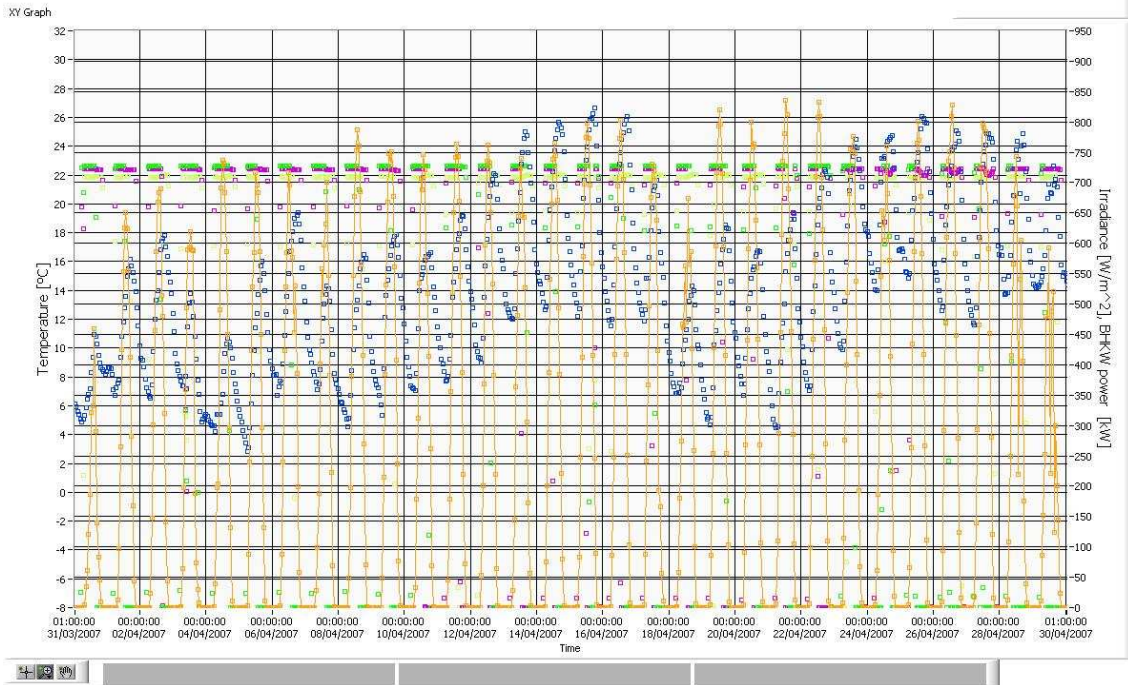


Figure 9 : Zoom sur les mesures du mois d'avril 2007 de la Figure 8.

A. 3 Glossaire

- AGID : Analyse et Gestion Intégrées et Durables des flux de matières et d'énergie en entreprise
 ASCII : American Standard Code for Information Interchange
 CRTE: Centre de Ressources des Technologies pour l'Environnement
 Enregistreur : appareil capable d'enregistrer des données pendant une durée déterminée.
 DCF77 : désignation d'un signal en ondes longues servant à la transmission du signal de l'horloge atomique du PTB Braunschweig.
 HNEC : Horaire normal d'Europe Centrale = UTC + 1h
 HAEC : Horaire avancé d'Europe centrale = UTC + 2h
 PTB : Physikalisch Technische Bundesanstalt Braunschweig
 UTC : Universal Time Coordinated (anglais), temps universel coordonné, précédemment désigné par GMT Greenwich Mean Time