

Vademekum Datenqualität und Datenaufbereitung für die Analyse

Inhaltsverzeichnis

1. Einleitung	3
2. Datenerfassung.....	3
2.1. Externe Datenquellen.....	3
2.2. Eigene Datenquellen	4
2.3. Absolute Zeitreferenz	4
2.4. Zeitstempel.....	5
2.5. Abtastrate f und Abtastintervall T	5
2.5.1. Erfassungszeitraum T	9
2.5.2. Zeitemstellung	9
3. Datenaufbereitung	10
3.1. Allgemeingültige Regeln.....	10
3.2. Bereinigung der Originaldatei	11
3.3. Format des Zeitstempel	12
3.4. Datenformat	13
3.5. Dateiname	13
3.6. Schlussfolgerung Datenaufbereitung	17
A. Anhang.....	18
A. 1 ASCII Tabelle	18
A. 2 Tool zur grafischen Darstellung von Messwerten	19
A. 3 Glossar	20

1. Einleitung

Im Rahmen des AGID¹ Projektes wurden mehrere Datenerhebungen oder auch Datenerfassungskampagnen durchgeführt, um Produktionsprozesse, Material- oder Energieflüsse zu analysieren und diese als Bezugsgrößen festzuhalten. Bei den Daten handelte es sich meist um Messwerte, die periodisch von Hand oder maschinell ausgelesen wurden und dann in beliebigen Formaten mit unbestimmter Datenorganisation und willkürlicher Benennung abgespeichert wurden. Dabei hat sich herausgestellt, dass die Qualität, Datenorganisation und Vollständigkeit der Datenserien in den wenigsten Fällen ausreichte, um sie ohne weitere zeitintensive Aufbereitung weiterverwenden zu können. Die Bearbeitung der Datenreihen nahm in einigen Fällen mehr Zeit in Anspruch als die nachfolgende Analyse. Deshalb wurde dieses Vademekum erstellt mit einer Vorgehensweise, nach welcher der Austausch und die Weiterverwendung von Daten erleichtert und die Qualität erhöht werden soll.

Da es keinen allgemeingültigen Standard für die Formatierung und die Datenorganisation gibt, wird hier ein Format vorgestellt, welches unabhängig von dem Betriebssystem und der Softwareanwendung von allen textverarbeitenden Maschinen gelesen werden kann. Durch eine konsequente Namensgebung wird eine rudimentäre Datenorganisation erreicht. Eine umfassendere Datenorganisation kann nachträglich durch den Import in anwenderspezifische Datenbanken durchgeführt werden.

2. Datenerfassung

Unter der Datenerfassung versteht man die systematische Sammlung und Speicherung von Informationen nach einem spezifischen Muster. Bei den im AGID Projekt untersuchten Produktionsprozessen, Material- und Energieflüssen handelt es sich um zeitabhängige Messwerte, welche Informationen über den Zustand von Teilprozessen enthalten.

2.1. Externe Datenquellen

Externe Datenquellen sind zur Verfügung gestellte Daten, die nicht beeinflussbar sind. Die Datenquellen können schon in Dateien oder Datenbanken archivierte Daten sein oder aber es wird ein Zugang zu bestehenden Datenerfassungsgeräten ermöglicht, von denen die Daten abgerufen werden können. Diese externen Datenquellen stellen bei der Datenerfassung oft die einzigen zugänglichen Informationen über den Betriebszustand eines Prozesses zur Verfügung. In den wenigsten Fällen wurden die Daten kontinuierlich gespeichert und archiviert, diese vom System mitgelieferte Option bleibt ungenutzt. Parameter die nicht in einem System zur Regelung benötigt werden, sind verständlicherweise für einen Anlagenbetreiber nicht interessant. Die Pflege und Wartung der Messtechnik und Parametrisierung bringt nach Meinung von Anlagenbetreibern keinen wirklichen Mehrwert für den Betrieb und wird entsprechend auch nicht durchgeführt. Dies liegt oftmals daran, dass Anlagen nach festeingestellten Regelgrößen betrieben werden, die iterativ durch Betriebserfahrung ermittelt wurden, aber so spezifisch sein können, dass auch eine angepasste Regelung keine befriedigenden Ergebnisse liefern kann (z.B. unbestimmbarer Zulauf (zeitl. und Zusammensetzung) im Pumpwerk eines Abwassersystems). Eine Datenspeicherung wird in den meisten Fällen als Nachweis für

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

Betriebsstörungen oder für die Lieferung von Energie- oder Stoffströmen geführt, d.h. im Normalfall wird sich niemand die archivierten Daten anschauen. Werden diese Daten zur Verfügung gestellt, so ist deren Format in den meisten Fällen vorgegeben und nicht von der Quelle her veränderbar. Es sollte in jedem Fall eine Plausibilitätsprüfung hinsichtlich der Aufzeichnungsperiode, des Datenerfassungsgerätes (eindeutige Identifikation) und dem zeitlichen Bezug zu einer absoluten Zeitreferenz durchgeführt werden. In Beispiel IV wird die Datenaufbereitung einer kontinuierlichen Fernabfrage des Energiezählers eines Energieversorgungsunternehmens (EVU) gezeigt.

2.2. Eigene Datenquellen

Unter eigenen Datenquellen versteht man jede Information über Prozesse, welche man selbst produziert hat. Dies können beispielsweise an der Anzeige eines Messgerätes abgelesene und manuell aufgezeichnete Werte sein oder automatisch gespeicherte Werte mittels Datenlogger, welcher nach eigenen Bedürfnissen parametrisiert und angepasst werden kann.

2.3. Absolute Zeitreferenz

Auch wenn nicht immer erforderlich, weil Prozesse zeitlich unabhängig voneinander laufen können, so ist der zeitliche Bezug der erfassten Daten zu einer absoluten Zeitreferenz angeraten. Die Zeit ist ein messtechnisch erfassbarer Parameter und wird vom menschlichen Bewusstsein als scheinbar kontinuierlich fortschreitende Ordnung von Ereignissen empfunden.² Daher ist die Kennzeichnung der Daten durch einen dazugehörigen Zeitstempel (siehe auch: Abschnitt 2.4 und Abschnitt 3.3) eine wichtige Option, wenn es um eine Betriebsanalyse oder die Erstellung von Zusammenhängen verschiedener Datensätze geht. Ein Bezug zu der absoluten Zeit kann zum Beispiel mittels GPS Uhr³ oder über einen aktiven DCF-77⁴ Empfänger hergestellt werden, welcher das von der PTB Braunschweig⁵ gesendete Zeitsignal empfangen kann. Handelt es sich um ein Messtechnik-Netzwerk, kann ein sogenannter interner oder externer NTP-Server⁶ eingesetzt werden, welcher die Zeitbasis für alle im Netzwerk verbundenen Geräte darstellt. Ein externer NTP Server stellt den Service, mit dem eine Verbindung durch das Internet hergestellt werden kann (z.B. Zeitserver der PTB Braunschweig: ptbtime1.ptb.de oder ptbtime2.ptb.de).

² Frei nach Wikipedia 2008

³ GPS Uhr: Synchronisierung der Uhr der Datenerfassung mittels GPS Empfänger (Global Positioning System)

⁴ DCF77 Langwellenempfänger: Empfänger, welcher das Funksignal der Atomuhrzeit an der PTB Braunschweig empfangen kann

⁵ PTB Braunschweig: Physikalisch Technische Bundesanstalt Braunschweig

⁶ NTP-Server: Network Time Protocol Server. Dieses Protokoll wurde mit dem Ziel entworfen, Rechner innerhalb lokaler Netzwerke und in Weitverkehrsnetzwerken zeitlich synchronisieren zu können. Das Protokoll basiert auf dem im Internet benutzten IP-Protokoll und ist für alle relevanten Betriebssysteme verfügbar (Quelle: PTB Braunschweig).

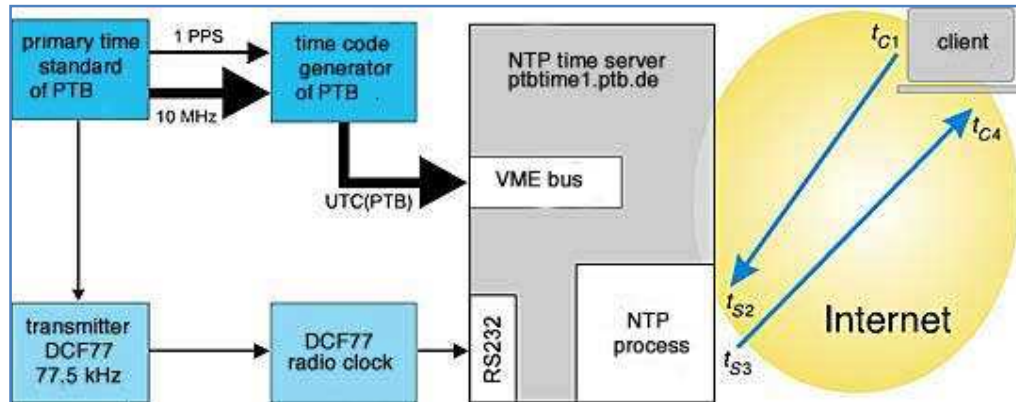


Abbildung 1: Funktionsschema einer NTP Server-Client Verbindung über das Internet zum Serverstandort PTB Braunschweig (Quelle: PTB Braunschweig)

2.4. Zeitstempel

Zu jedem Messwert oder sonstigen Information sollte ein zeitlicher Bezug hergestellt werden können. Dies wird mit einem sogenannten Zeitstempel realisiert. Damit wird ermöglicht, betriebsübergreifende Zusammenhänge zu erkennen, welche von verschiedenen Datenerfassungssystemen sowie von örtlich entfernten Datenquellen stammen können. Diese Daten können Störgrößen (wie z.B. meteorologische Ereignisse auf eine Gebäudeklimatisierung) oder ein Potenzial und einen Bedarf an Stoff- und Energieströmen von benachbarten Systemen darstellen, welche eine Abstimmung mit dem zu betrachtenden System erlauben könnten oder Kosten und Preise von Stoffen und Energien, welche Alternativen zu den bisher verwendeten Betriebsmitteln darstellen könnten, etc. Mit dem Zeitstempel wird also eine Synchronisation der Daten ermöglicht, d.h. es kann ein zeitlicher Bezug der Daten zueinander hergestellt werden.

2.5. Abtastrate f und Abtastintervall T

Die Häufigkeit von periodischen Messungen pro Zeitintervall wird **Abtastrate** (auch Abtastfrequenz) genannt (Formelzeichen f , Einheit Hertz [Hz]). Das Zeitintervall zwischen zwei Messpunkten wird **Abtastintervall** genannt (Formelzeichen T ⁷, Einheit Sekunde [s]). Bei Abtastintervallen unter einer Sekunde wird meist die Abtastrate angegeben.⁸ Es wird in der Datenerfassung zwischen der Abtastrate für die eigentliche Messung und der Abtastrate für die Speicherung der Messwerte unterschieden. Werden Kurzzeitmessungen durchgeführt, so können diese beiden Abtastraten identisch sein, da jeder Messwert einzeln abgespeichert wird. Bei Langzeitmessungen werden meist die Messwerte mit einer hohen Abtastrate gemessen und innerhalb des Abtastintervalls für die Datenspeicherung zwischengespeichert. Gespeichert wird dann nur ein Wert, welcher aus den zwischengespeicherten Werten rechnerisch ermittelt wurde (arithmetisches Mittel, gleitendes Mittel, Integral, Differential, etc.)

⁷ Das Zeitintervall T ist der Kehrwert der Abtastrate bzw. -frequenz: $T=1/f$, $f=1/T$.

⁸ Beispiel: eine Abtastrate von 10 [Hz] entspricht einem Abtastintervall von $T=1/(10 \text{ [Hz]}) = 0.1 \text{ [s]} = 100 \text{ [ms]}$

Mindestabtastrate f_{\min}

Off ist ein Messgerät schon in Betrieb, bevor eine Messkampagne durchgeführt wird bzw. Daten aus der Vergangenheit betrachtet werden. Insofern besteht keine Möglichkeit, die Abtastrate optimal einzustellen. Kann jedoch innerhalb der Planung oder Installation die Abtastrate eines Messsystems beeinflusst werden, so sollte die Auslegung bzw. die Einstellung der Abtastrate nach dem Nyquist-Shannon⁹ Theorem durchgeführt werden:

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

Dieses Theorem besagt, dass die Abtastrate mindestens zweimal so groß sein muss wie die Frequenz des Signals, das gemessen wird. So wird sichergestellt, dass alle dynamischen Effekte des zu untersuchenden Systems erfasst werden bzw., dass keine Verfälschung der Dynamik des Messsignals angezeigt wird (z.B. Schwebung, Auslöschung).

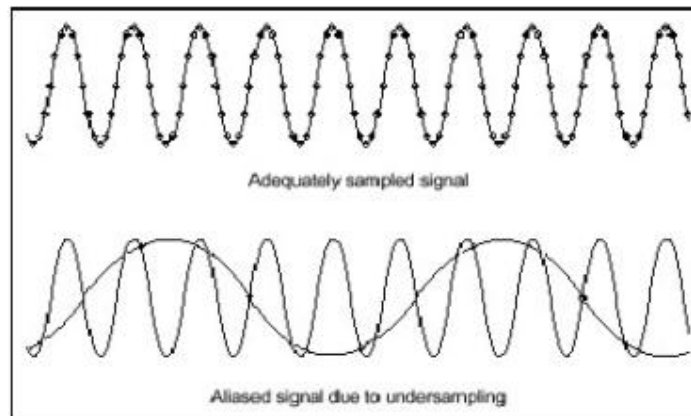


Abbildung 2: Messpunkte der oberen Kurve mit korrekt eingestellter Abtastrate; Messpunkte der unteren Kurve mit zu niedrig eingestellter Abtastrate (engl.:undersampling): Eine Schwebung wird überlagert dargestellt, wie sie vom Messgerät anhand der Messpunkte und der Funktion für die Ermittlung einer harmonische sinusförmigen Schwingung ausgegeben wird. Daraus kann fälschlich auf eine 5-fach kleinere Frequenz des Signals geschlossen werden. (Quelle: National Instruments).

Maximale Abtastrate f_{\max}

Die maximal mögliche Abtastrate eines Messgerätes gibt an, mit welcher Frequenz Messwerte ausgegeben werden können. Hierbei ist die Abtastrate der Messwertausgabe maßgebend. Bei integrierenden Messsystemen wie beispielsweise Energiezählern, ist die Abtastrate der eigentlichen Messung relativ hoch (im Kilohertzbereich bei digitalen, quasi unendlich bei analogen Geräten¹⁰). Die Abtastrate für die Messwertausgabe ist jedoch um ein vielfaches geringer. Das Abtastintervall der Datenspeicherung mittels Datenlogger kann wiederum bei Bedarf an die Abtastrate der Messwertausgabe angepasst werden. Dieses Abtastintervall ist üblicherweise relativ groß (eine Viertelstunde oder eine Stunde). Die Auflösung¹¹ S der Anzeige oder Ausgabeschnittstelle des Messgerätes ergibt sich aus der Anzahl n von Impulsen oder Umdrehungen einer Zehlscheibe pro Zählleinheit. Das

⁹ Harry Nyquist, Physiker und Claude Elwood Shannon, Mathematiker

¹⁰ „Quasi unendlich“: limitierende Faktoren in Bezug auf „unendlich“: die Masse der sensorischen Bauteile und die Reibung von mechanischen Lagerungen, d.h. minimale Änderungen wirken zwar auf das Messsystem ein, können aber nicht erfasst werden

¹¹ Auflösung: kleinste feststellbare Veränderung eines Messwertes

minimale Abtastintervall für die Speicherung von Messwerten eines solchen Messgerätes¹² kann ermittelt werden, indem die Auflösung mit der maximalen Anschlussleistung multipliziert und dann der Kehrwert gebildet wird. Wenn gewährleistet werden soll, dass alle gezählten Ereignisse erfasst (gespeichert) werden sollen, so muss der Datenlogger auf eine minimale Abtastrate eingestellt werden.

I. Beispiel zum Abgleich der Mindestabtastrate der Messwerterfassung und Ermittlung des minimal möglichen Abtastintervalls für die Datenspeicherung

In diesem Beispiel wird anhand von technischen Spezifikationen einer Datenerfassung eines Energiezählers für einen elektrischen Verbraucher erläutert, wie die Abtastrate der Messwerterfassung und die Dynamik des zu messenden Signals aufeinander abgestimmt werden müssen. Danach wird dargestellt mit welchem minimalen Abtastintervall es noch Sinn macht, die Daten zu speichern, um alle Informationen über die Veränderung von Messwerten erfassen zu können. Die Datenerfassung ist wie folgt aufgebaut: ein Energiezähler mit Impulsschnittstelle ist in einem Stromkreis zwischen 230 VAC Anschluss und elektrischem Verbraucher installiert. Ein Impulszähler ist an den Impulsausgang des Energiezählers angeschlossen. Dieser summiert alle Impulse auf, die der Energiezähler liefert und speichert dann diesen Wert. An den Impulszähler ist ein Datenlogger angeschlossen, der diesen Wert periodisch abfragt, in einen Energiemesswert umwandelt und mit einem Zeitstempel versehen abspeichert.

Technische Daten des Energiezählers:

Maximale Anschlussleistung: $P_{Max} = 24 \text{ [kW]}$

Auflösung der Impulsschnittstelle: $S = 1'000 \text{ Impulse/[kWh]} = 1 \text{ Impuls/[Wh]}$

Minimale Impulsdauer: $T_{Impulse} = 30 \text{ [ms]} = 0.03 \text{ [s]}$

Zunächst muss ermittelt werden, ob der Impulszähler in der Lage ist, alle Impulse erfassen zu können. Dazu wird die minimale Abtastrate für den Impulszähler errechnet.

Impulsfrequenz aus der minimalen Impulsdauer:

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

Minimale Abtastrate des Impulszählers f_{scan} für die Erfassung aller Impulse am Impulsausgang:

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

Maximales Abtastintervall für die Erfassung aller Impulse am Impulsausgang:

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

d.h. bei der Auswahl eines Impulszählers muss darauf geachtet werden, dass dieser eine Mindestabtastrate von 66.67 [Hz] gewährleistet, damit alle Impulse des Energiezählers gezählt werden können. Bei jedem erkannten Impuls wird der Zahlenwert einer natürlichen Zahl in einem nichtflüchtigen Speicher des Impulszählers um eine Einheit aufsummiert. Der Zahlenwert im nichtflüchtigen Speicher wird auch bei einer Unterbrechung der Versorgungsspannung erhalten bleiben. Die gezählte Energie des Energiezählers kann

¹² Datenspeicherung z.B. realisiert durch eine Kombination aus Energiezähler, der Impulse ausgibt, einen Impulszähler, der diese Impulse zählt und aufsummiert, und einen Datenlogger, der diese aufsummierten Impulse in Energiemesswerte umrechnet und speichert.

durch Division des gespeicherten Zahlenwertes mit der Auflösung der Impulsschnittstelle S ermittelt werden.

Die maximale Abtastrate (oder das minimale Abtastintervall) für die Speicherung der gezählten Energie, wenn die Auflösung der Impulsschnittstelle bei maximaler Anschlussleistung voll ausgenutzt werden soll, d.h. jede kleinste Änderung gespeichert werden soll:

maximale Abtastrate pro Impulse bei maximaler Leistung:

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

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

minimales Abtastintervall: $T_{min} = 1/f_{Max}$

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

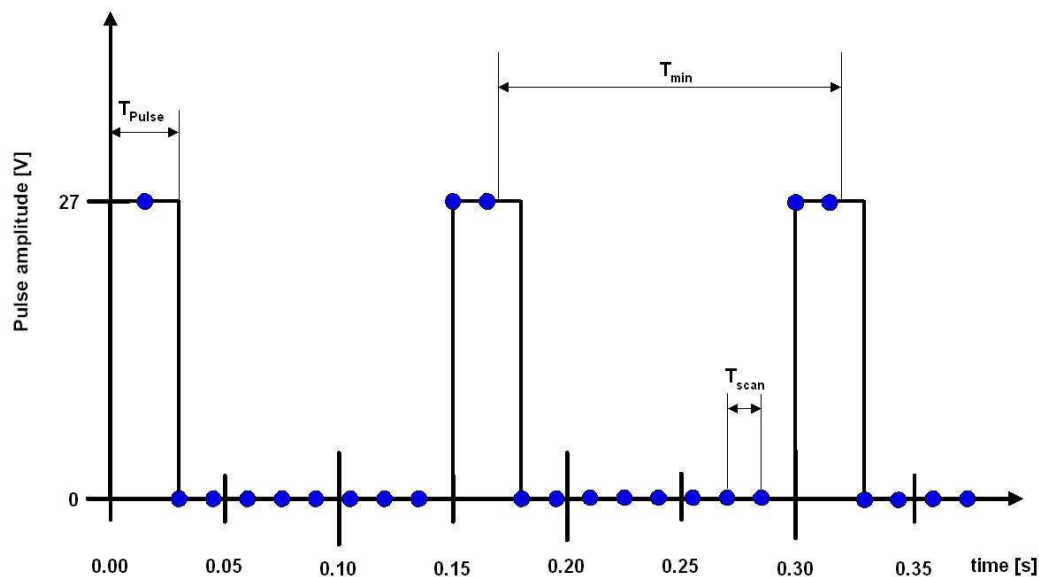


Abbildung 3: Darstellung der im Beispiel errechneten Abtastintervalle für die Datenerfassung eines Energiezählers mit S0 Schnittstelle unter maximaler Anschlussleistung. Die Impulsdauer ist konstant mit $T_{pulse} = 30$ Millisekunden angegeben. Das maximale Abtastintervall des Impulzzählers (blaue Punkte), der an der S0 Schnittstelle eines Energiezählers angeschlossen ist $T_{scan} = 15$ Millisekunden und das minimale Abtastintervall für die Datenspeicherung eines Datenloggers, der am Ausgang des Impulzzählers angeschlossen ist, ist $T_{min} = 150$ Millisekunden (0.15 Sekunden).

Optimale Abtastrate

Die Ermittlung einer optimalen Abtastrate in einem Datenerfassungssystem führt meist zu einer Kompromisslösung. Mit Einstellung der höchstmöglichen Abtastrate erhält man die detaillierteste Information über das dynamische Verhalten eines Systems. Es ist aber klar, dass bei der Speicherung jedes einzelnen Wertes mit einer sehr hohen Abtastrate auch die größten Speicherkapazitäten schnell aufgebraucht sind. Deshalb ist es wichtig, die Dynamik eines Systems zu kennen und zu überlegen, ob, wann und in welcher Auflösung diese Informationen benötigt werden, um bei der Datenerfassung den zur Verfügung stehenden Speicherplatz effektiv zu nutzen und die Datenverarbeitung zu erleichtern. Die

maximale Abtastrate für die Datenspeicherung aus dem Zahlenbeispiel in Beispiel I macht keinen Sinn, wenn in einer speziellen Anwendung nur ein Bruchteil der maximalen Anschlussleistung gemessen wird oder sich die gemessene Leistung zeitlich nur langsam ändert. Grundsätzlich sind kleine Abtastintervalle auch für länger andauernde Messkampagnen sowie Monitoringsysteme sinnvoll, wenn neben Betriebszuständen auch eine unbekannte Dynamik des Systems, d.h. unbekannte zeitliche Änderungen der Messwerte untersucht werden sollen. Das Abtastintervall für die Speicherung der Messwerte über einen längeren Erfassungszeitraum liegt erfahrungsgemäß zwischen einer Minute und bis zu einer Stunde.

Ein nachträglicher Abgleich der Abtastraten verschiedener Messdatenreihen ist machbar, falls erforderlich. Dies ist der Fall, wenn für eine Analyse oder Modellierung eine gemeinsame Abtastrate für alle Messdatenreihen – im folgenden „Zeitschrittzeilen“ genannt - benötigt wird. Es muss die größte gemeinsame Abtastrate aller Zeitschrittzeilen ermittelt werden. Folgende Überlegungen sind bei der nachträglichen Verringerung der Abtastrate anzustellen: Bei integrierenden Messgrößen wie z.B. Energie oder Füllstand, ist der Zeitpunkt des letzten Messpunktes im Zeitstempel zu übernehmen. Bei Messgrößen, die Momentanwerte darstellen wie z.B. Durchfluss, Leistung, Temperatur, etc. wird bei der Verringerung der Abtastrate ein Mittelwert gebildet. Dementsprechend liegt der dazugehörige Zeitpunkt auch in der Mitte zwischen dem ersten und letzten Erfassungszeitpunkt. Wenn nun beide Messgrößen in Zusammenhang gebracht werden, so gilt für beide der gemeinsame Zeitpunkt am Ende der gemeinsamen Abtastrate.

II. Beispiel: Synchronisierung von Zeitschrittzeilen mit unterschiedlichen Abtastintervallen

Die Zeitschrittzeile einer Energieerfassung wurde mit einem Abtastintervall von 1 Minute abgespeichert. Diese soll nun mit einer anderen Zeitschrittzeile mit einem Abtastintervall von 1 Stunde synchronisiert werden. Die Minutenwerte werden dabei von Beginn bis Ende je einer Stunde aufsummiert und skaliert (Summe über 60 Werte pro Stunde, Skalierung: 60 Wattminuten entspricht einer Wattstunde) und mit einem neuen Zeitstempel versehen. Wenn die Zeitstempel beider Zeitschrittzeilen nun übereinstimmen, können die Messwerte auf einen gemeinsamen Zeitstempel bezogen werden und sind somit zeitlich synchronisiert.

2.5.1. Erfassungszeitraum T

Unter Erfassungszeitraum versteht man den Zeitraum, in dem kontinuierlich Daten gespeichert wurden. Vorhandene Lücken in diesem Zeitraum z.B. durch einen Ausfall der Datenerfassung, können durch Extrapolation oder wie im Falle der meteorologischen Daten durch Abgleich mit vorhandenen Daten der nächstgelegenen meteorologischen Messstation (beispielsweise am Airport Findel, Luxembourg) aufgefüllt werden. In der Jahressumme werden kleine Ungenauigkeiten in der Lückenbeseitigung zwar nicht erkennbar sein, aber eine lückenlose Datensammlung bereitet in der nachfolgenden Bearbeitung, Analyse und Modellierung weniger Probleme.

2.5.2. Zeitumstellung

Liegen im Erfassungszeitraum ein oder mehrere Tage an denen eine Zeitumstellung von Sommer- auf Winterzeit und/oder umgekehrt erfolgt, so müssen die dabei entstehenden

Lücken oder Überlagerungen beseitigt werden. Auf unserem Längengrad empfiehlt es sich als Zeitbasis die UTC¹³ Zeit oder die UTC + 1[h] einzustellen. Letztere entspricht unserer mitteleuropäischen Winterzeit. Auf jeden Fall soll ein automatischer Wechsel zwischen Sommer- und Winterzeit vermieden werden, um die Bearbeitung der Datenreihen erheblich zu erleichtern: Bereinigung des „Datenverlustes“ von einer Stunde bei der Umstellung von Winter- auf Sommerzeit durch den „Zeitsprung“ von 2:00 Uhr auf 3:00 Uhr und die Überlagerung der Daten bei der Umstellung von Sommer- auf Winterzeit durch das Zurücksetzen von 3:00 Uhr auf 2:00 Uhr.

3. Datenaufbereitung

In diesem Abschnitt wird ein Datenformat vorgestellt, welches klar strukturiert und einfach in der Weiterverarbeitung ist. Desweiteren wird anhand eines Beispiels die Bearbeitung eines weniger optimalen, spezifischen Datenformats zu einem allgemein gut verwendbaren Format gezeigt. Dies soll als Denkanstoß für den Aufbau eines neuen oder die Änderung eines bestehenden Datenerfassungssystems dienen. Am CRP Henri Tudor liegt nach mehreren Jahren Tätigkeit im Bereich Messen, Steuern und Regeln die Erfahrung vor, dass die Daten aus externen Datenquellen immer wieder in nicht einfach in allgemein verwendbarer Form vorliegen. Für eine einzelne Anwendung oder ein spezifisches System mag dies nicht von Nachteil sein, wenn aber systemübergreifend Daten zusammen geführt werden sollen, so ist es von Vorteil, wenn einige allgemeingültige Regeln beachtet werden. Einen internationalen Standard, der außerdem allen Anforderungen zur Datenverarbeitung gerecht wird, gibt es nach dem heutigen Kenntnisstand leider noch nicht. Die Aufbereitung der Rohdatensätze oder besser gesagt der Originaldaten ist für jeden Datensatz unterschiedlich aufwändig. Eine Aufbereitung muss in den meisten Fällen immer dann durchgeführt werden, wenn Daten aus unterschiedlichen Quellen synchronisiert werden müssen, um diese dann zusammen analysieren zu können.

3.1. Allgemeingültige Regeln

Alle Daten werden im ASCII Format in der Form tabulatorgetrennte¹⁴ Textdatei mit der Dateiendung „.txt“ abgespeichert. Dies hat den Vorteil, dass die Daten auch in Zukunft noch mit einer beliebigen textverarbeitenden Software gelesen werden können ohne an eine bestimmte Anwendung gebunden zu sein, die es dann eventuell nicht mehr geben wird. Alle Zusatzinformationen zu den Messwerten (sogenannte Metadaten) werden in die Kopfzeilen (Header) des jeweiligen Datensatzes geschrieben. Das erleichtert die Suche nach Maßeinheiten, Gerätespezifikationen und bietet beim Import der Datensätze in andere Softwareanwendungen eine einfache Möglichkeit, den Daten Eigenschaften zuzuordnen (vor allem die richtige Zuordnung von Maßeinheiten). Dieser Header soll durch ein eindeutiges und unverwechselbares Zeichen vom eigentlichen Dateninhalt getrennt werden, das üblicherweise nicht im Dateninhalt zu finden ist. Es sollte kein Zeichen verwendet werden, das als Platzhalter eingesetzt werden könnte oder in einem üblichen Texteditor nicht sichtbar dargestellt werden kann, wie beispielsweise ein

¹³ UTC = universal time coordinated (engl.) bzw. koordinierte Weltzeit (deutsch), früher GMT Greenwich mean time oder mittlerer Greenwichzeit

¹⁴ In der Informatik wird die Bezeichnung „durch Tabulatorzeichen getrennt“ häufig in englischer Sprache verwendet: tab separated.

Zeilenumbruch. Am besten ist die multiple Verwendung eines Zeichens, um Verwechslungen mit Zeichen zu vermeiden, die in einer Programmiersprache eine andere Bedeutung haben können.

Das Raute-Zeichen¹⁵ # oder der Asterisk * als übliche Kommentarzeichen können so ### oder so **** oder eine eindeutige Textbezeichnung als Trennzeichen für die Kopfzeile eingesetzt werden.

III. Beispiel: Aussehen einer typischen Kopfzeile (Header) eines Datenfiles

Inhalt einer Kopfzeile eines Messdatenfiles:

```
***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 Pyrheliometer [°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]

```
#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]
```

Das Trennzeichen von Kopfzeile und Dateninhalt ist hierbei die gelb markierte Textpassage: #data table header# Der nachfolgende Text wird als Bezeichner für die einzelnen Datenspalten eingesetzt.

3.2. Bereinigung der Originaldatei

Bevor eine Manipulation an Originaldaten vorgenommen wird, sollte immer eine Sicherungskopie angelegt werden. Jede Veränderung der Originaldaten sollte dokumentiert werden, damit eventuell auftretende Fehler identifiziert und beseitigt werden können. Außerdem kann diese Dokumentation dem Urheber der Originaldaten dazu

¹⁵In der Informatik gebräuchliche englische Bezeichnung: hash

dienen, künftige Daten in der gewünschten Form bereitzustellen. Der Abschnitt der Textdatei, welcher die eigentlichen Daten beinhaltet, darf keine multiplen Zeichen (insbesondere keine nicht-druckbaren Zeichen wie Tabulator, Zeilenumbruch und Zeilenvorschub) zwischen den Daten enthalten. Diese können beim Import der Daten in ein Tabellenformat die Position der Daten ungewollt und eventuell auch unbemerkt verändern. Da das ASCII Format nur eine begrenzte Anzahl von Schriftzeichen hat, können viele Sonderzeichen und Schriftzeichen nicht dargestellt werden. Siehe Tabelle im Anhang.

3.3. Format des Zeitstempel

Das Format des Zeitstempels wird für alle Datenreihen wie folgt angegeben (nach ISO 8601:2004 und EN 28601):

YYYY-MM-DD\shh:mm:ss

YYYY: Jahr in vier Ziffern z.B. 2007

MM: Monat in zwei Ziffern z.B. Februar = 02

DD: Tag des Monats in zwei Ziffern z.B. 09

\s: Leerzeichen (informatisch engl. „Backslash s“), ASCII Zeichen dezimal Nummer 32

hh: Stunde in zwei Ziffern von 00 bis 23

mm: Minute in zwei Ziffern von 00 bis 59

ss: Sekunde in zwei Ziffern von 00 bis 59

Mit diesem Zeitformat ist sichergestellt, dass beim Import der Daten im ASCII Format in eine Analysesoftware oder in eine Datenbank (z.B. LabVIEW, Matlab, MS Office, etc.) die zeitliche Zuordnung eindeutig ist. Ein anderes Format für den Zeitstempel wäre die sogenannte Unix-Zeit, die im 32-bit Integer Format angegeben wird (in Sekunden seit dem 1. Januar 1970 00:00 UTC). Aus den folgenden zwei Gründen sollte dieses Zeitformat aber nicht gewählt werden: 1) Die wenigsten Menschen werden beim Betrachten einer ganzen Zahl diese ohne Probleme einem Datum und einer Uhrzeit zuordnen können. 2) Die 32-bit Integer Zahl wird am 19. Januar 2038 ihr Maximum erreichen und dann überschrieben werden. Man rechnet hier wie bei dem Jahr-2000-Bug mit einem ähnlich gelagerten Problem.

Auch wenn mit der Verwendung des hier vorgestellten Formats für den Zeitstempel die Probleme bei der Erkennung in den am meisten genutzten Softwareanwendungen beseitigt sein dürften, so wird angeraten, auch den Inhalt des Zeitstempels kritisch zu betrachten. Dies gilt insbesondere für den Datumswechsel und die letzte und erste Zeitangabe des Tages. In manchen Datensätzen endet der Tag mit 00:00:00 Uhr oder mit 24:00:00. Beide Uhrzeiten werden beim Import in eine Software nach Zuweisung des Zeitformates sehr wahrscheinlich falsch interpretiert und somit wird ein falscher Zeitstempel erzeugt. Besonders kurios ist der sogenannte „Gastag“ wie er von Unternehmen in der Gaswirtschaft verwendet wird. Hier beginnt ein neuer Tag um 06:00:00 Uhr und endet um 05:59:59 am darauffolgenden Tag. Damit keine falschen Interpretationen durch die Software auftreten können, sind die folgenden Uhrzeiten für den Anfang und das Ende

eines Tages wie folgt festgelegt: Beginn eines neuen Tages um 00:00:00 und Ende des Tages um 23:59:59.

3.4. Datenformat

Für die eigentlichen Daten/ Messwerte wird der Punkt als Dezimaltrennzeichen festgelegt. Genau drei Nachkommastellen sind nach Möglichkeit zu vermeiden, damit diese nicht mit der im deutschsprachigen Gebrauch häufig verwendeten Tausendertrennung verwechselt werden kann (z.B.: die drei Nachkommastellen der Zahl 123.456 aufgerundet zu 123.46 oder eine zusätzliche „0“ angehängen zu 123.4560). Die regionalen Einstellungen des Computers müssen vor dem Datenimport daraufhin geprüft werden. Eine Tausendertrennung (Beispiel: 1000 wird zu 1'000) ist zwar sehr gut für die Lesbarkeit einer großen Zahl, wird aber hier nicht verwendet, um Fehler bei der Interpretation nach einem Datenaustausch mit Computern zu vermeiden, die unterschiedliche regionale Einstellung besitzen. Die so aufbereiteten Messwerte werden ohne Maßeinheit in einer Zeile hinter den Zeitstempel geschrieben. Das Feldtrennzeichen zwischen dem Zeitstempel und dem Messwert ist ein Tabulator [\t] (ASCII Zeichen dezimal 09, Bezeichnung TAB, Symbol \t); nach jeder fertigen Zeile erfolgt ein Zeilenumbruch, bestehend aus zwei Zeichen [\r\n]¹⁶: Wagenrücklauf (ASCII Zeichen dezimal 13, Bezeichnung CR, Symbol \r) und Zeilenvorschub (ASCII Zeichen dezimal 10, Bezeichnung LF, Symbol \n).

Beispiel für eine Messwertreihe (Tabulator, Leerzeichen und Zeilenumbruch sind nicht sichtbar dargestellt):

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

Mit diesem Format werden in den meisten Software-Anwendungen, die auf Tabellenkalkulation oder Textverarbeitung basieren, die Daten in einer 2-dimensionalen Tabelle dargestellt und können so leicht weiterverarbeitet werden. Des weiteren ist zu beachten, dass Tabellenkalkulationsprogramme oft auf eine maximale Zeilenzahl von $2^{16} = 65'536$ und eine maximale Spaltenzahl von $2^8 = 256$ begrenzt sind. Sollen mehr Messwerte in einem Datensatz verarbeitet werden, so wird empfohlen, diese in eine Datenbank zu importieren.

3.5. Dateiname

Bei Dateien, die Messwerte enthalten, ist es immer von Vorteil, wenn aus dem Dateinamen entnommen werden kann, um welche Messwerte es sich handelt und wann und von wem diese aufgezeichnet wurden. Um eine historische Ordnung¹⁷ in dem Datenverzeichnis zu

¹⁶ \r\n: carriage return line feed, zu Deutsch: Wagenrücklauf Zeilenvorschub ist im windowsbasiertem Textformat als Zeichen für den Zeilenumbruch festgelegt.

¹⁷ Historische Ordnung: Es ist schon viel darüber diskutiert worden, welche Art der Ordnung die optimalste sei (alphabetische oder zeitbasierte). Da sich die Messwerte meist auf einen Zeitraum beziehen, hat sich diese Art der Ordnung als die praktikabelste erwiesen. Die Dateinamen nach alphabetischer Ordnung können innerhalb einer Benennung auch zeitbasiert sortiert werden, wenn das Datum angehängen wird.

erhalten, in welche diese Dateien gespeichert werden, wird folgendes Aussehen für den Dateinamen vorgeschlagen:

Datum_Uhrzeit_Messobjekt_Operateur.XXX

Datum: YYYYMMDD

Uhrzeit: hhmmss (kann nach eigenem Ermessen auch ausgelassen werden)

Messobjekt: Kurzbezeichnung oder Akronym der Messstelle oder des Sensors, wie z.B. Kesseltemperatur als Kesseltemp oder Meteorologische Messstation Findel als MeteoFindel

Operateur: Person, Firma oder Organisation, die für die Datenaufzeichnung verantwortlich ist.

.XXX: Dateinamenserweiterung¹⁸. Ein Punkt wird vor diese als Abtrennung vom übrigen Dateinamen gesetzt (z.B. .txt oder .xls).

IV. Beispiel einer Datenaufbereitung

In diesem Abschnitt wird anhand eines praktischen Beispiels schrittweise dargestellt, wie Originaldaten in ein Format gebracht wurden, das weiterverarbeitet werden kann. Die Daten wurden von der Stadt Esch/Alzette zur Verfügung gestellt. Das Format ist auf den ersten Blick übersichtlich und innerhalb eines Tabellenkalkulationsprogramms kann damit gearbeitet werden. Aber zur Herstellung einer zeitlichen Synchronisation zu anderen Daten, welche parallel an anderen Messstellen im gleichen System aufgezeichnet wurden, ist dieses Format so nicht zu gebrauchen. Das Ausgangsformat ist in Bild 1 verkürzt dargestellt. Es handelt sich um eine 2-dimensionale Tabelle. Die Spalten werden einer Uhrzeit, die Zeilen einem Datum zugeordnet. Es findet eine Unterbrechung der kontinuierlichen Darstellung zum Datum des 28.10.2007 statt, an welchem die Uhrzeit von Sommer- auf Winterzeit umgestellt wurde. Die letzte Uhrzeit eines Tages wird mit 00:00 Uhr angegeben. Ziel dieses Beispiels ist es, die Darstellung einer Methode wie diese für die weitere Bearbeitung nicht optimale 2-dimensionale Tabelle, in eine serielle Zeitreihe mit einem korrekten Zeitstempel pro Zeile und dem dazugehörigen Messwert umzuwandeln.

1. Die Daten werden von der Stadt Esch/Alzette monatlich per Email zugeschickt. Dabei ist die Benennung der Dateien jeden Monat gleich, d.h. die Dateien müssen umbenannt werden, um das Überschreiben zu verhindern, aber vor allem, um die spätere Identifikation aus einer großen Datensammlung zu erleichtern. Die Originaldateinamen werden in einem Emailarchiv beibehalten. In diesem Beispiel wird die Datei „Generator1.csv“ in „200710_BHKW_1_.csv“ umbenannt.
2. Zunächst wird geprüft, ob in einem monatlichen Datensatz eine Zeitumstellung von mitteleuropäischer Zeit (MEZ) auf mitteleuropäische Sommerzeit (MESZ) stattgefunden hat (Siehe Bild 1). Dies wird im Datensatz in einer gesonderten Zeile vermerkt. Diese Daten müssen zum Vergleich mit anderen Zeitreihen in das gemeinsame MEZ Zeitformat gebracht werden (und vor allem deshalb, um Lücken oder überlagerte Messreihen während der Zeitformatänderung zu vermeiden).

¹⁸ informatisches Englisch: filename extension

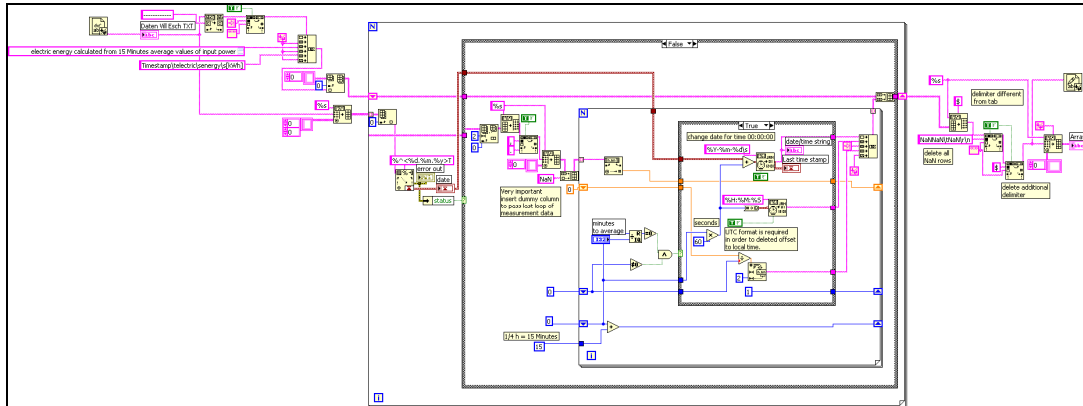


Bild 2: Algorithmus zur Bearbeitung der Rohdaten des EVU Stadt Esch/Alzette in der grafischen Programmiersprache LabVIEW

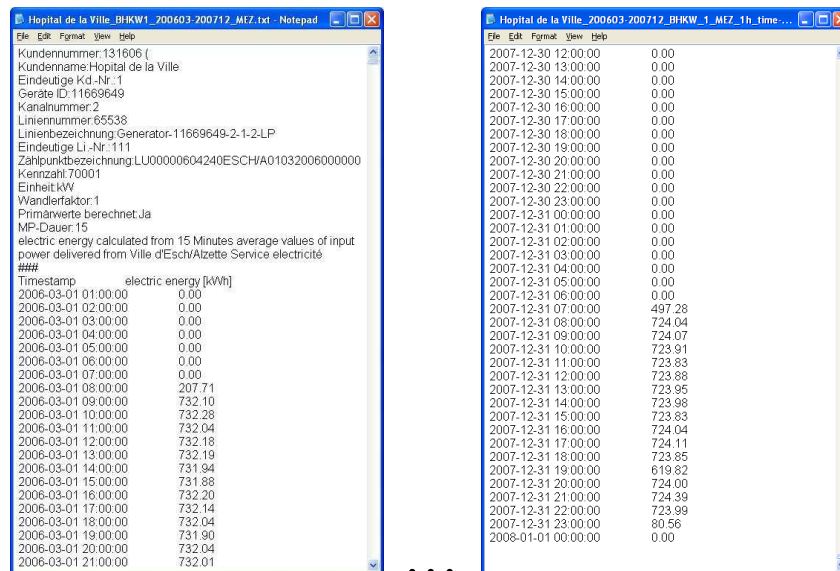


Bild 3: Fertige Zeitschrittdatei nach der Datenaufbereitung: Lastkurve des BHKW 1 Generatorleistung vom Oktober 2007 mit einheitlichem Zeitstempel in MEZ

- Der Inhalt aller einzelnen Arbeitsdateien wird in eine Gesamtdati kopiert und für die weitere Analyse als fertige Zeitschrittdatei abgespeichert (siehe Bild 3). Mit Hilfe dieser Zeitschrittdatei können alle Messwerte in einer Tabelle in Spalten nebeneinander dargestellt, verglichen und analysiert werden.

3.6. Schlussfolgerung Datenaufbereitung

Es hat sich während der Bearbeitung der unterschiedlichen Teilprojekte im Projekt AGID gezeigt, dass bei der Datenerfassung die Aufbereitung vorhandener Messwerte und Verbrauchsdaten immens zeitaufwendig sein kann. Ähnliche Erfahrungen werden auch in anderen Projekten gemacht. Es sollte daher bei der Planung eines neuen Systems für eine Datenerfassung oder bei der Definition einer Datenerfassungskampagne immer darauf geachtet werden, dass die Aufbereitungszeit schon durch entsprechende Konfiguration und Abstimmung des Systems erheblich minimiert werden kann. Wo dies bei schon bestehenden Systemen nicht möglich ist, können die in diesem Vademekum dargestellte Vorgehensweise und die Beispiele als Orientierungshilfe dienen. Es wird angestrebt, mit den in künftigen Projekten gesammelten Erkenntnissen und Erfahrungen, dieses Vademekum kontinuierlich zu aktualisieren.

A. Anhang

A.1 ASCII Tabelle

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

Abbildung 4: ASCII Tabelle. Die Zeichen 0 – 31 und 127 sind sogenannte Steuerzeichen, die nicht druckbar sind.

A. 2 Tool zur grafischen Darstellung von Messwerten

Für eine erste visuelle Untersuchung der Daten im Projekt AGID wurde ein LabVIEW Programm erstellt, mit welchem alle Messwerte aus Zeitschrittfolgen, die nach dem im Abschnitt 3 beschriebenen Format aufgebaut sind, grafisch dargestellt werden können. Diese grafische Darstellung bietet Zoomfunktionen, Ausblendung, Achsenverschiebung, etc. zur visuellen Analyse der Zeitschrittfolgen (siehe Bild 4 und Bild 5). Ein Executable dieses Programms kann bei Bedarf beim CRTE angefragt werden.

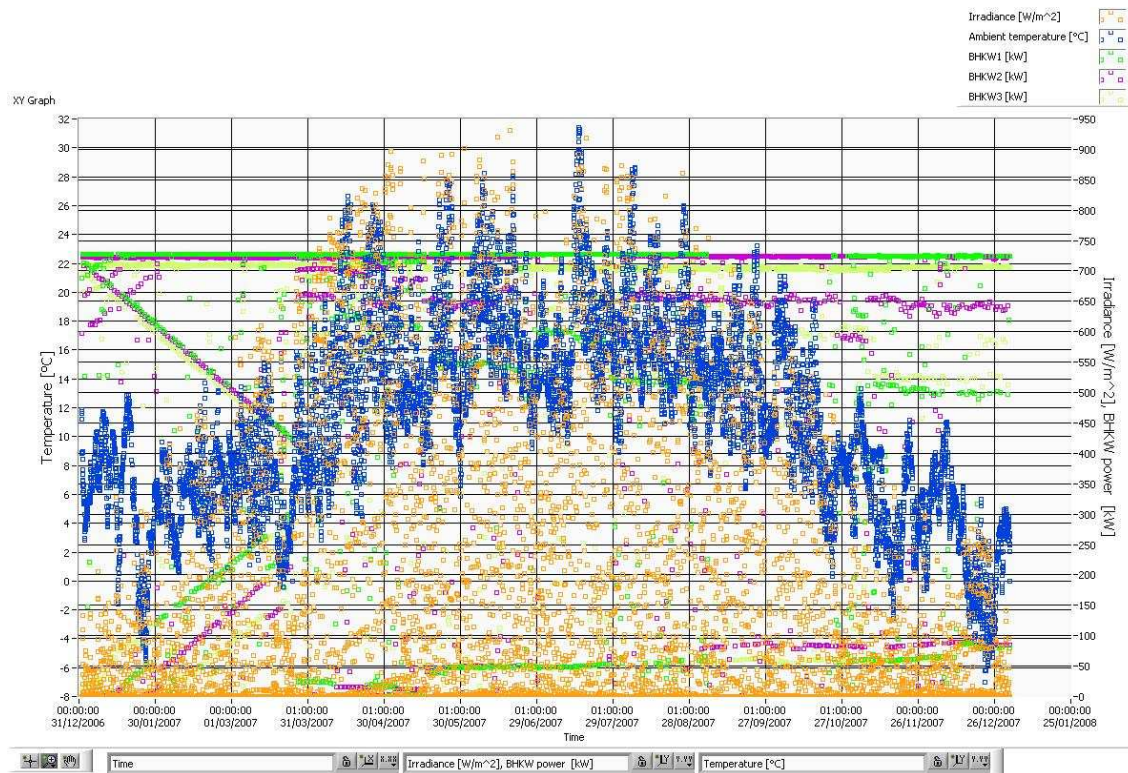


Bild 4: Alle bisher vorhandenen kontinuierlich aufgezeichneten Messwerte zusammen in der Jahresübersicht mit einem Abtastintervall von einer Stunde (1 [h]).

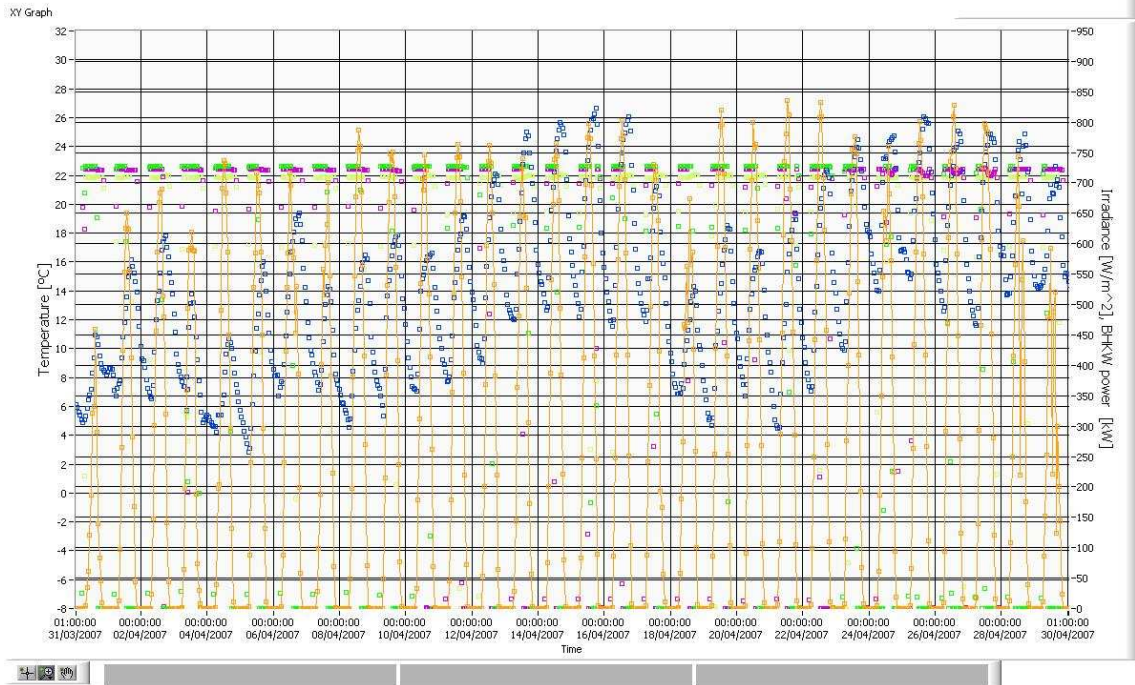


Bild 5: Ein Zoom im Diagramm in Bild 4 auf die Messwerte des Monats April 2007.

A. 3 Glossar

- AGID: Analyse et Gestion Intégrées et Durables des flux de matières et d'énergie en entreprise
- ASCII: American Standard Code for Information Interchange
- BHKW: Blockheizkraftwerk. Kombiniertes Heizkraftwerk mit Generator zur Erzeugung von elektrischem Strom
- CRTE: Centre de Ressources des Technologies pour l'Environnement
- Datenlogger: Gerät, welches Daten über einen bestimmten Zeitraum speichern kann.
- DCF77: Bezeichnung eines Langwellenfrequenzsignals, mittels der das Atomuhrsignal der PTB Braunschweig übertragen wird.
- EVU: Energieversorgungsunternehmen
- MEZ: Mitteleuropäische Zeit = UTC + 1h
- MESZ: Mitteleuropäische Sommerzeit = UTC + 2h
- PTB: Physikalisch Technische Bundesanstalt Braunschweig
- UTC: Universal Time Coordinated (engl.), koordinierte Weltzeit (deutsch), früher GMT Greenwich Mean Time oder mittlerer Greenwichzeit