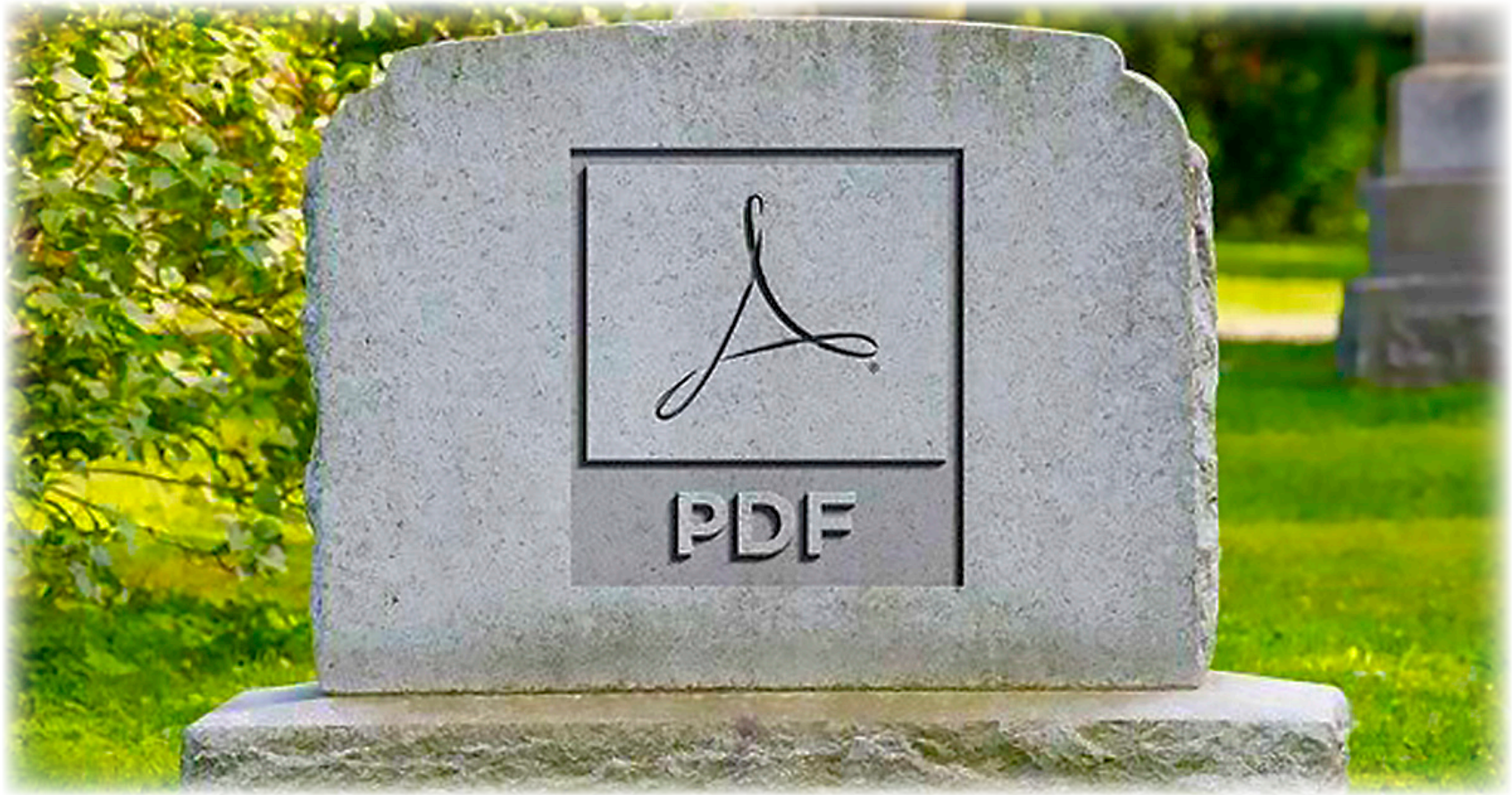
PDF & Open Data:
PDF Beyond Final Form Visual Content

Dov Isaacs
Principal Scientist, Adobe Systems Incorporated

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

A PDF Association Presentation · © 2018 by PDF Association · www.pdfa.org

The Challenge – Combating Ignorance & Misconceptions



PDFs are where data goes to die!

Adobe

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

The Challenge – Combating Ignorance & Misconceptions

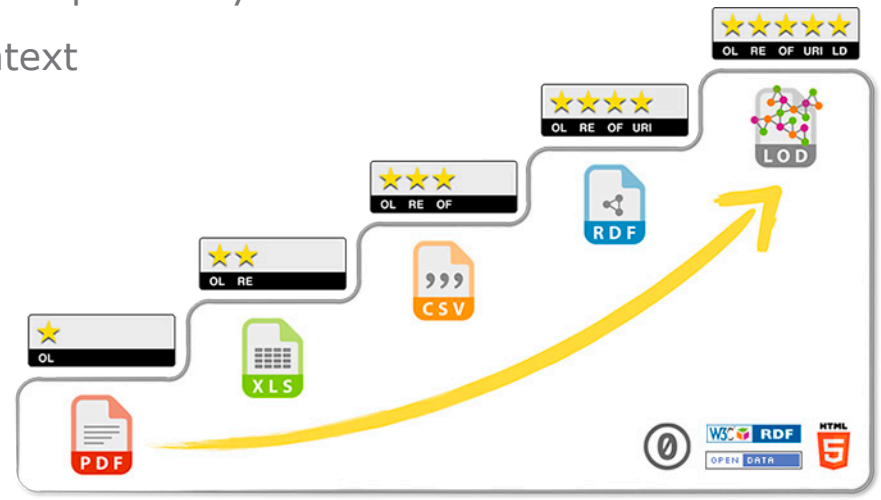
- **The "PDFs are where data goes to die" mantra is mindlessly echoed around the internet**
  - **Great *clickbait* for attracting on-line attention**
  - **Most often parroted by persons with "issues" both real and imagined**
    - Experienced difficulty in getting data requested, both in terms of availability and method of delivery
    - Undue attention to Tim Berners-Lee's 5-star ranking system for Open Data in which he gives PDF 1 out of 5 stars
    - User confusion & "know it alls" who don't
    - The PDF *versus* "Web Formats" bogeyman
    - Mistaken beliefs that PDF is a proprietary file format; Irrational hatred of PDF and/or Adobe

Adobe

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

The Challenge – Combating Ignorance & Misconceptions

- **Tim Berners-Lee's 5-star ranking system for Open Data (2010)**
  - ★ Make your stuff available on the Web (whatever format) under an open license
  - ★★ Make it available as structured data (e.g., Excel instead of image scan of a table)
  - ★★★ Make it available in a non-proprietary open format (e.g., CSV instead of Excel)
  - ★★★★ Use URIs to denote things, so that people can point at your stuff
  - ★★★★★ Link your data to other data to provide context



  - **Ignores reality that PDF may contain data files; it's not just for scraping**
  - **Questionable value judgments; XLS may be more appropriate than CSV (hint: he hates Microsoft?)**
- **Our challenge is to tease apart the issues and provide data access solutions appropriate to one's needs**

Adobe

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

Open Data

- The concept of *Open Data* is that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.
- Some contexts in which Open Data is of vital importance:
  - Functioning of governance in a free society;
    laws supporting data access
    (e.g., United States Freedom of Information Act and similar laws worldwide)
  - STEM Research;
    access to full underlying data of experiments
    is necessary for result validation and peer review
  - Finance;
    certain financial data, especially of public companies
    fits under the category of Open Data

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

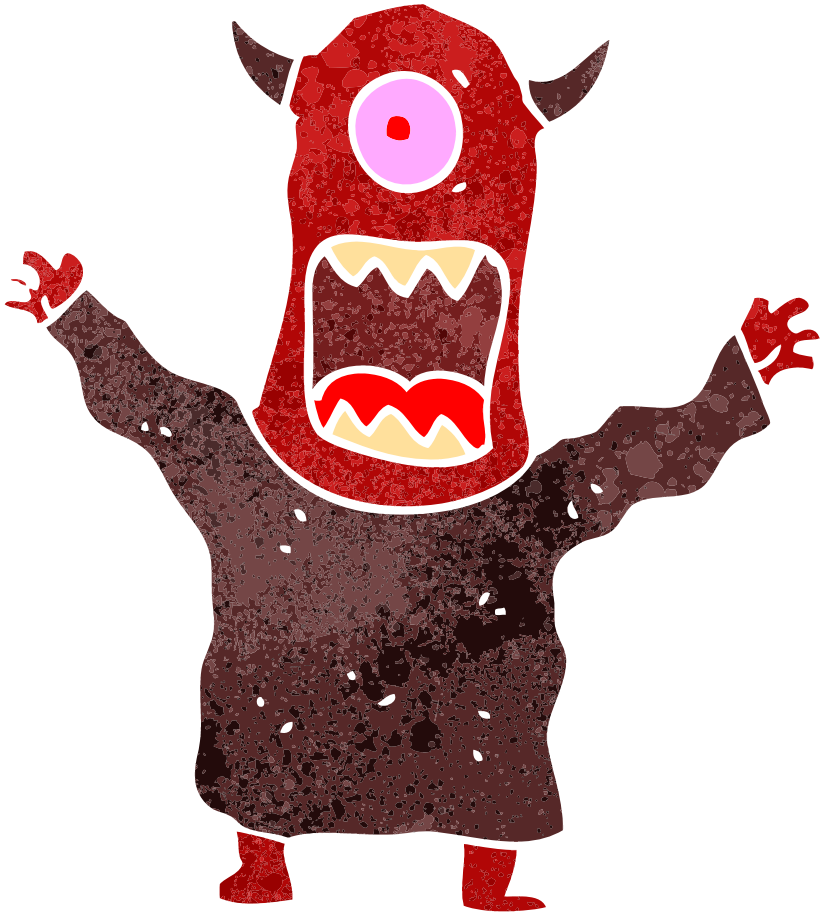
Open Data

- **Issues**
  - **Which data qualifies for Open Data?**
  - **Intentional delays in releasing data (i.e., roadblocks)**
  - **Release of data in inappropriate formats requiring onerous and error-prone OCR, export, and "scraping" processes**
    - Database or spreadsheet data output to a PDF file
    - Statistical data or even text as a PDF file created by scanning n^{th} generation photocopies
    - Recipient needs "source document" and not "final form file format" for republishing content
  - **Recipient doesn't know how to use data provided**
  - **Recipient wants / needs more than entitled to**
  - **Dynamic access versus static, blind exchange archival access**

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

The PDF versus "Web Formats" Bogeyman

- **Web Formats (HTML, CSS, Javascript, et.al.)**
  - **Potential to provide great *dynamic* access to data**
  - **Great for dashboard applications**
  - **Requires on-line access and availability of all assets**
  - **Results may vary by browser (HTML & CSS are not ISO standards, but rather, W3C "recommendations")**
  - **May change underneath you!**
  - **Not archival!**

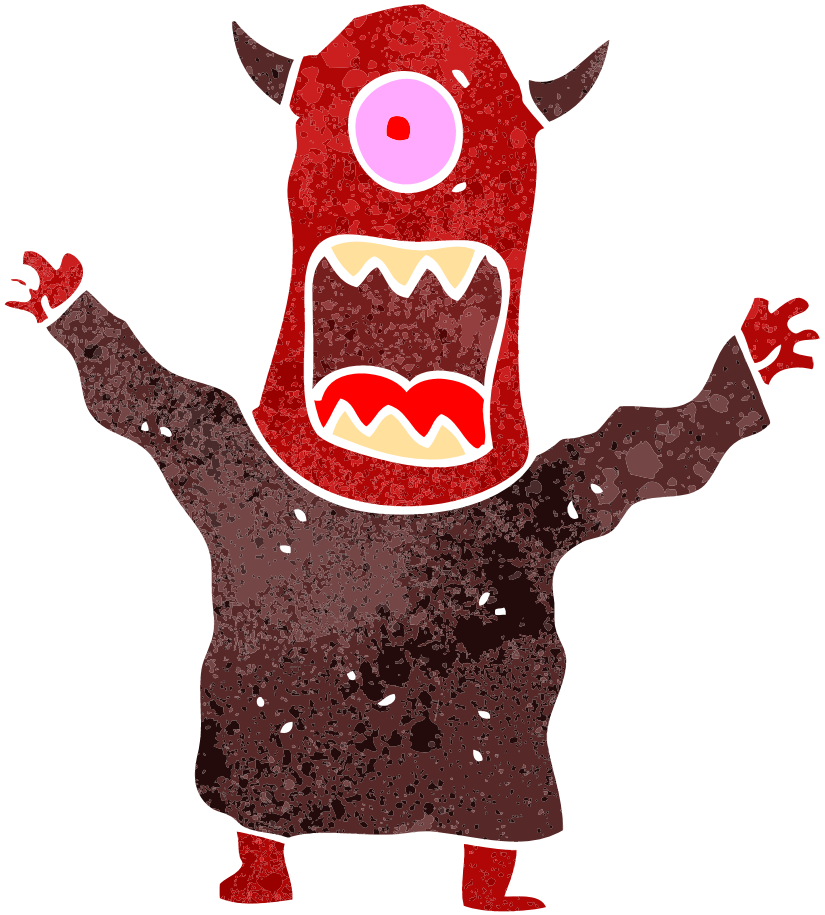


Adobe

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

The PDF versus "Web Formats" Bogeyman

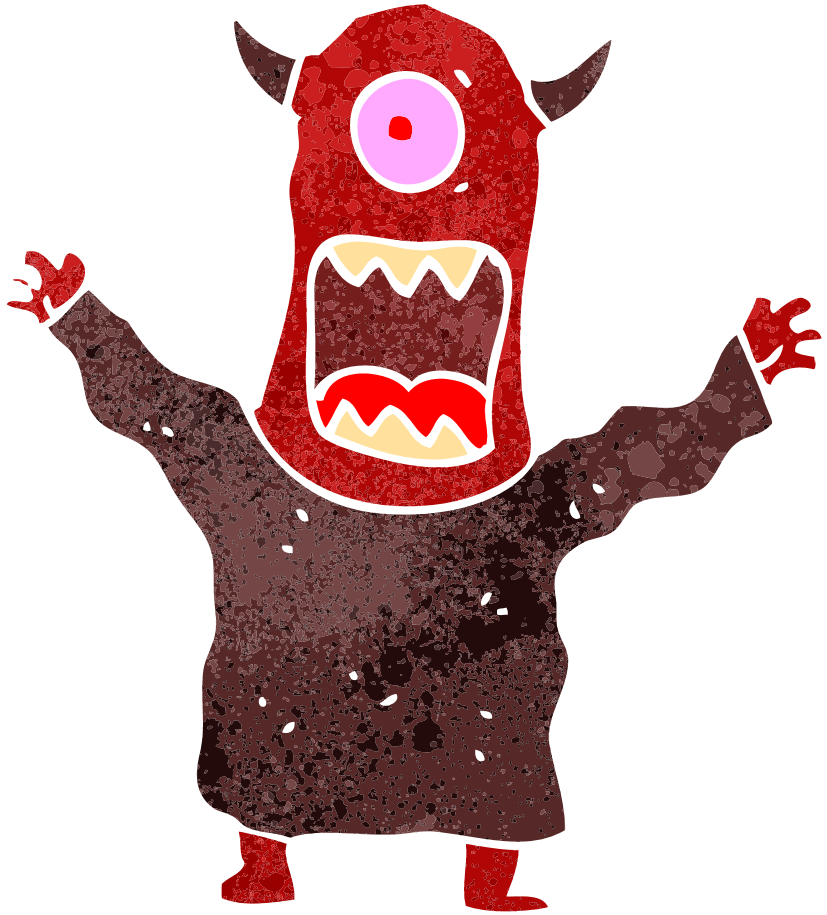
- **PDF**
  - **A "container" file format primarily used as a *final form file format* supporting compressed file attachments and metadata**
  - **May be fully self-contained ("blind exchange")**
  - **Potential to provide great *static* access to data (offline and doesn't change)**
  - **International standards:**
    - ISO 32000-1, PDF 1.7 & ISO 32000-1, PDF 2.0
    - ISO 19005, PDF/A (for archiving) subset standards



Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

The PDF versus “Web Formats” Bogeyman

- **Hybrid Solutions (Web Formats & PDF)**
  - **Initial access of data-laden PDF file via the internet**
    - Perfectly reasonable
    - Once you’ve downloaded the PDF file, you are self-sufficient
  - **PDF file accesses external data via hyperlinks**
    - Not recommended
    - Defeats the purpose of using PDF



Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

Using PDF for (Open) Data

- **Available techniques and approaches**
  - **Embedded files**
    - Easier to implement
    - Independent of PDF viewer (no “context issues”); data extracted and operated on by program associated with particular file type
  - **Metadata & Tagging**
    - Generally requires more complex programming
    - Requires PDF viewer that is cognizant of particular metadata / tagging usage to access underlying data and pass on to program that should process that data

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

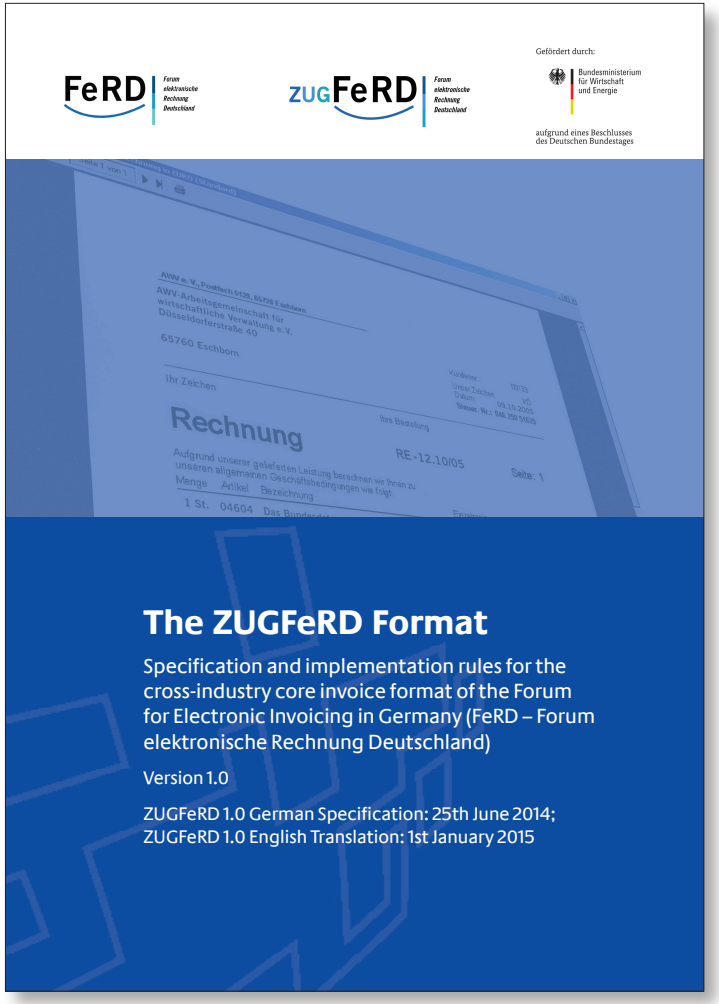
Using PDF for (Open) Data

- **Best practices for use of PDF for *any* data, much less Open Data:**
  - **Use PDF/A-3**
  - **Embed data files using PDF/A-3 *associated files* feature in conjunction with *attachment annotations* (providing direct access to data from within PDF graphical content)**
  - **Use "standard formats" for embedded data files wherever possible or practical**
  - **Do not embed "executable" files**
  - **Check files for viruses and exploits *prior* to embedding**
  - **Do not use hyperlinks to reference data external to the PDF file**
  - **Digitally sign PDF file to assure no modifications are made**

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

A Simple Example – ZUGFeRD

- Specification and implementation rules for the cross-industry core invoice format of the Forum for Electronic Invoicing in Germany (FeRD - Forum elektronische Rechnung Deutschland)
- Combines a *Visual Representation* (ISO 19005-3-2012 PDF/A-3) with a XML-based *Structured Representation* (ISO 15000-5:2014, the UN/CEFACT Cross Industry Invoice [CII] Standard plus the European Core Invoice Data Model MUG) of the invoice



Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

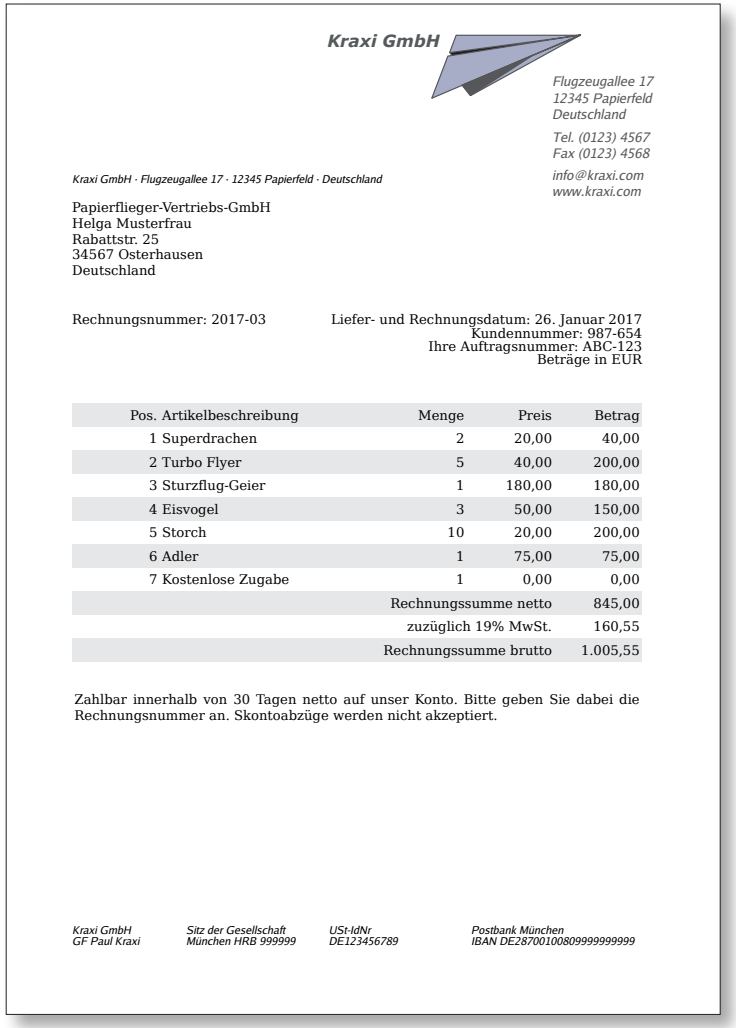
A Simple Example – ZUGFeRD

- PDF/A-3 is the format used for ZUGFeRD
  - It is able to meet all the requirements made of an electronic invoice in a single file.
  - It is able to embed all data used for automated further processing.
  - Invoicees who do not carry out any automatic further processing are not irritated by additional files.
  - PDF is already the most widely used data format for exchanging documents. (Invoices can be viewed using the free tools which come preinstalled on virtually all PCs, smart phones and other end devices.)
  - PDF/A ensures that an invoice file can be viewed permanently in a form that remains identical.
- Other supporting documents/data may also be attached
- 1 PDF/A-3 file = 1 invoice

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

A Simple Example – ZUGFeRD

- **A sample ZUGFeRD File**



Kraxi GmbH

Flugzeugallee 17
12345 Papierfeld
Deutschland

Tel. (0123) 4567
Fax (0123) 4568

info@kraxi.com
www.kraxi.com

Kraxi GmbH · Flugzeugallee 17 · 12345 Papierfeld · Deutschland

Papierflieger-Vertriebs-GmbH
Helga Musterfrau
Rabattstr. 25
34567 Osterhausen
Deutschland

Rechnungsnummer: 2017-03

Liefer- und Rechnungsdatum: 26. Januar 2017
Kundennummer: 987-654
Ihre Auftragsnummer: ABC-123
Beträge in EUR

Pos.	Artikelbeschreibung	Menge	Preis	Betrag
1	Superdrachen	2	20,00	40,00
2	Turbo Flyer	5	40,00	200,00
3	Sturzflug-Geier	1	180,00	180,00
4	Eisvogel	3	50,00	150,00
5	Storch	10	20,00	200,00
6	Adler	1	75,00	75,00
7	Kostenlose Zugabe	1	0,00	0,00
			Rechnungssumme netto	845,00
			zuzüglich 19% MwSt.	160,55
			Rechnungssumme brutto	1.005,55

Zahlbar innerhalb von 30 Tagen netto auf unser Konto. Bitte geben Sie dabei die Rechnungsnummer an. Skontoabzüge werden nicht akzeptiert.

Kraxi GmbH
GF Paul Kraxi

Sitz der Gesellschaft
München HRB 999999

USt-IdNr
DE123456789

Postbank München
IBAN DE28700100809999999999

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

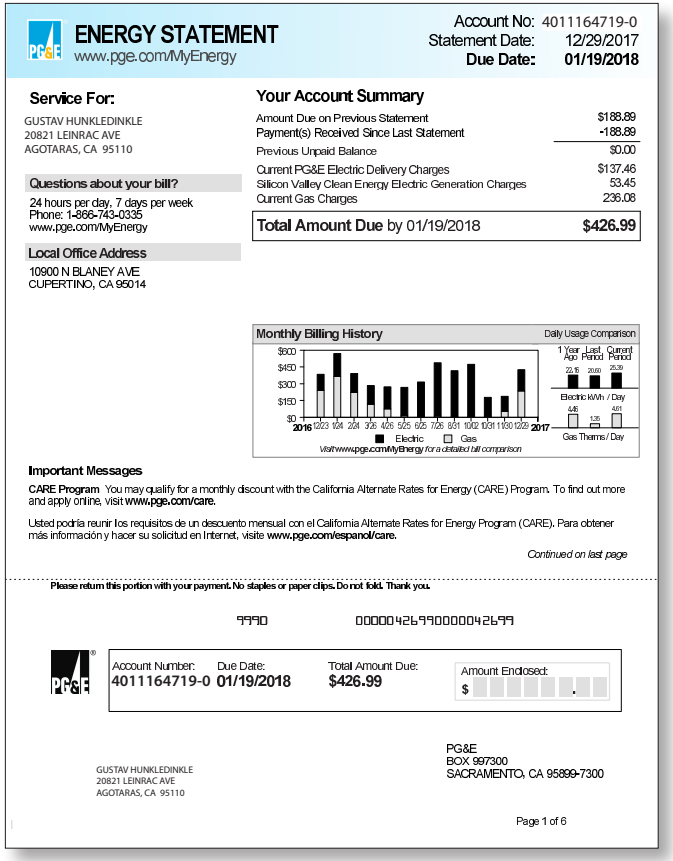
Consumer Utility Statements

Pacific Gas & Electric Company (PG&E)

- Customers receive monthly printed statements. The envelope also includes advertisements and regulatory information (notices of rate increases usually) on separate sheets plus a return envelope.
- Customers may also download a PDF copy of the statement from the company's website for a limited period of time.
- The statement contains current billing information plus some limited historical information in graphical form.
- If you know where to look, you can also download historical data associated with your account and other useful information from *elsewhere* on the PG&E website!

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

Consumer Utility Statements



ENERGY STATEMENT
www.pge.com/MyEnergy

Account No: 4011164719-0
Statement Date: 12/29/2017
Due Date: 01/19/2018

Service For:
GUSTAV HUNKLEDINKLE
20821 LEINRAC AVE
AGOTARAS, CA 95110

Questions about your bill?
24 hours per day, 7 days per week
Phone: 1-866-743-0335
www.pge.com/MyEnergy

Local Office Address
10900 N BLANEY AVE
CUPERTINO, CA 95014

Your Account Summary

Amount Due on Previous Statement	$188.89
Payment(s) Received Since Last Statement	-188.89
Previous Unpaid Balance	$0.00
Current PG&E Electric Delivery Charges	$137.46
Silicon Valley Clean Energy Electric Generation Charges	53.45
Current Gas Charges	236.08
Total Amount Due by 01/19/2018	**$426.99**

Monthly Billing History — Daily Usage Comparison

Visit www.pge.com/MyEnergy for a detailed bill comparison

Important Messages

CARE Program You may qualify for a monthly discount with the California Alternate Rates for Energy (CARE) Program. To find out more and apply online, visit **www.pge.com/care.**

Usted podría reunir los requisitos de un descuento mensual con el California Alternate Rates for Energy Program (CARE). Para obtener más información y hacer su solicitud en Internet, visite **www.pge.com/espanol/care.**

Continued on last page

Please return this portion with your payment. No staples or paper clips. Do not fold. Thank you.

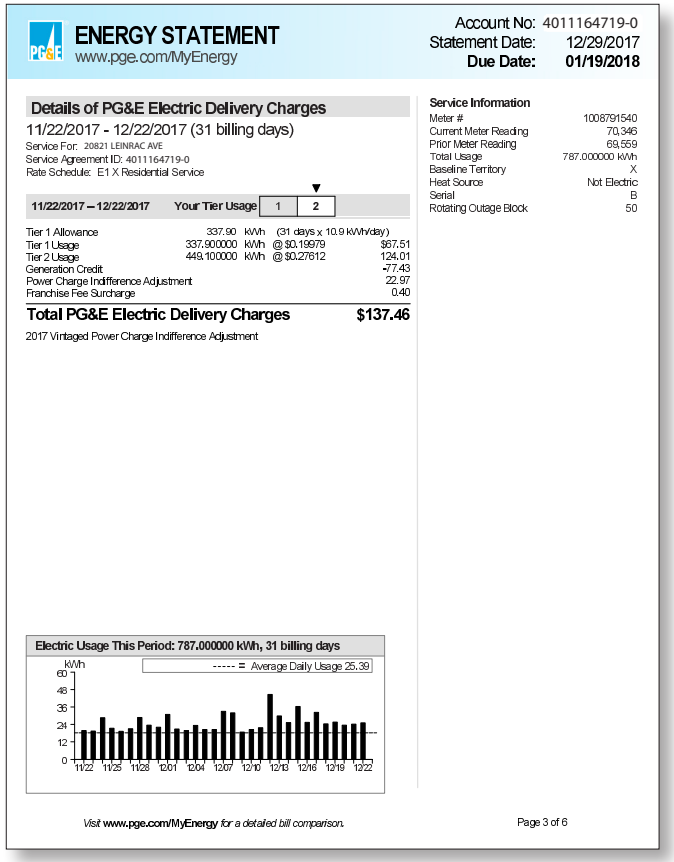
9990 0000042699000042699

Account Number: 4011164719-0 Due Date: 01/19/2018 Total Amount Due: $426.99 Amount Enclosed: $

PG&E
BOX 997300
SACRAMENTO, CA 95899-7300

GUSTAV HUNKLEDINKLE
20821 LEINRAC AVE
AGOTARAS, CA 95110

Page 1 of 6



ENERGY STATEMENT
www.pge.com/MyEnergy

Account No: 4011164719-0
Statement Date: 12/29/2017
Due Date: 01/19/2018

Details of PG&E Electric Delivery Charges

11/22/2017 - 12/22/2017 (31 billing days)

Service For: 20821 LEINRAC AVE
Service Agreement ID: 4011164719-0
Rate Schedule: E1 X Residential Service

11/22/2017 – 12/22/2017 Your Tier Usage: 1 2

Tier 1 Allowance	337.90 kWh	(31 days x 10.9 kWh/day)	
Tier 1 Usage	337.900000 kWh	@ $0.19979	$67.51
Tier 2 Usage	449.100000 kWh	@ $0.27612	124.01
Generation Credit			-77.43
Power Charge Indifference Adjustment			22.97
Franchise Fee Surcharge			0.40
Total PG&E Electric Delivery Charges			**$137.46**

2017 Vintaged Power Charge Indifference Adjustment

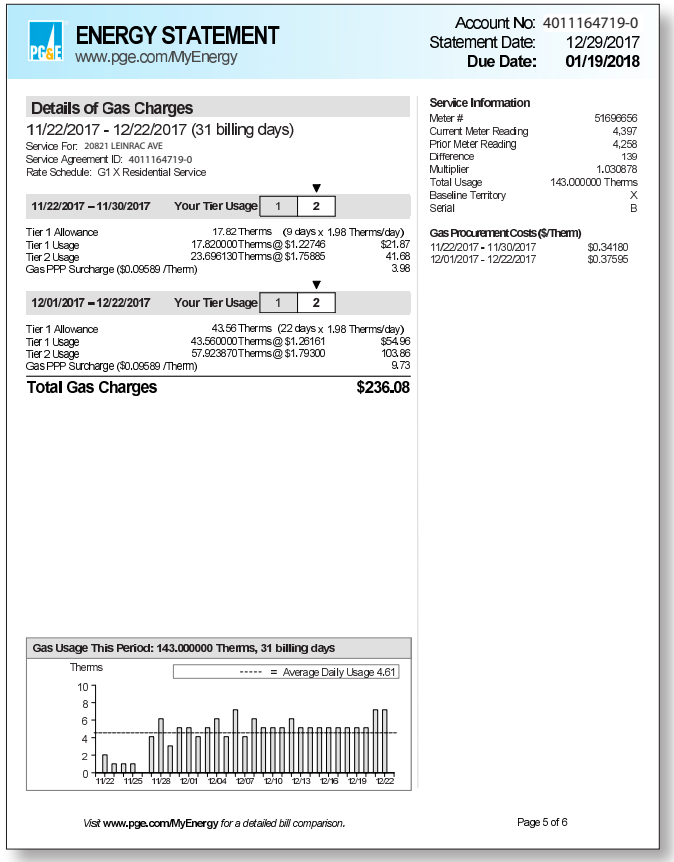
Service Information

Meter #	1008791540
Current Meter Reading	70,346
Prior Meter Reading	69,559
Total Usage	787.000000 kWh
Baseline Territory	X
Heat Source	Not Electric
Serial	B
Rotating Outage Block	50

Electric Usage This Period: 787.000000 kWh, 31 billing days

Visit www.pge.com/MyEnergy for a detailed bill comparison.

Page 3 of 6



ENERGY STATEMENT
www.pge.com/MyEnergy

Account No: 4011164719-0
Statement Date: 12/29/2017
Due Date: 01/19/2018

Details of Gas Charges

11/22/2017 - 12/22/2017 (31 billing days)

Service For: 20821 LEINRAC AVE
Service Agreement ID: 4011164719-0
Rate Schedule: G1 X Residential Service

11/22/2017 – 11/30/2017 Your Tier Usage: 1 2

Tier 1 Allowance	17.82 Therms	(9 days x 1.98 Therms/day)	
Tier 1 Usage	17.820000 Therms	@ $1.22746	$21.87
Tier 2 Usage	23.696130 Therms	@ $1.75885	41.68
Gas PPP Surcharge ($0.09589 /Therm)			3.98

12/01/2017 – 12/22/2017 Your Tier Usage: 1 2

Tier 1 Allowance	43.56 Therms	(22 days x 1.98 Therms/day)	
Tier 1 Usage	43.560000 Therms	@ $1.26161	$54.96
Tier 2 Usage	57.923870 Therms	@ $1.79300	103.86
Gas PPP Surcharge ($0.09589 /Therm)			9.73
Total Gas Charges			**$236.08**

Service Information

Meter #	51699656
Current Meter Reading	4,397
Prior Meter Reading	4,258
Difference	139
Multiplier	1.030878
Total Usage	143.000000 Therms
Baseline Territory	X
Serial	B

Gas Procurement Costs ($/Therm)

11/22/2017 - 11/30/2017	$0.34180
12/01/2017 - 12/22/2017	$0.37595

Gas Usage This Period: 143.000000 Therms, 31 billing days

Visit www.pge.com/MyEnergy for a detailed bill comparison.

Page 5 of 6

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

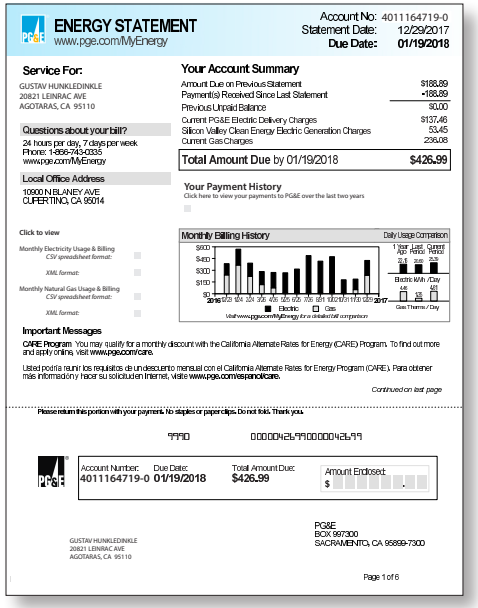
Consumer Utility Statements

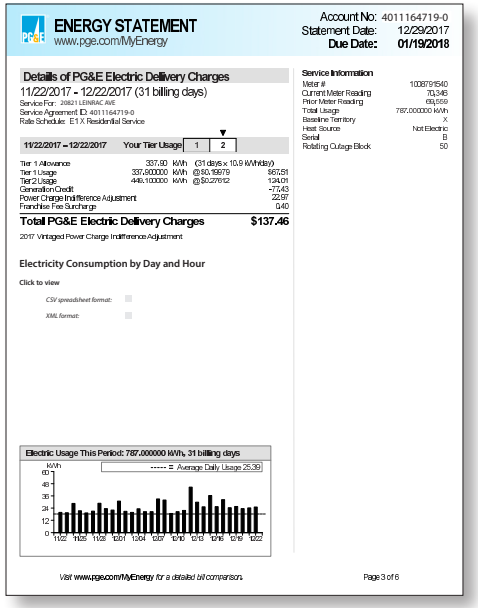
- **Stage 1 Fixups:**
  - **Downloaded all the "goodies" from the PG&E website**
  - **Simply attach these files to the base PDF file?**
    Don't waste your time!
    Not very useful!
  - **Embedded the downloaded files via *attachment annotations* at appropriate locations within the statement's PDF file**
  - **Created a PDF/A-3b file via Acrobat Pro DC Preflight profile *Convert to PDF/A-3b (without fallback conversion)***

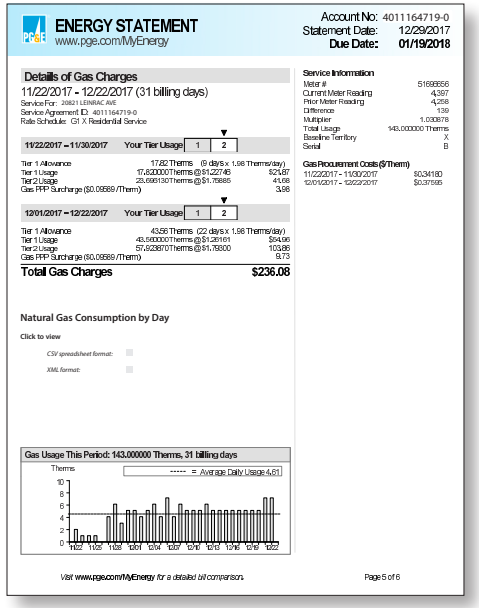
Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

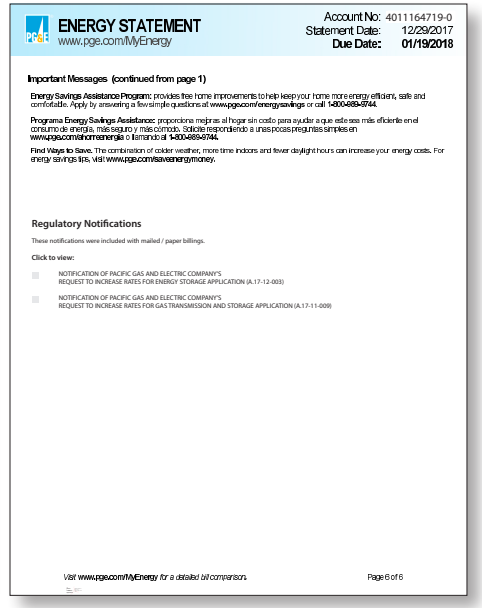
Consumer Utility Statements

- **Stage 2 Fixups:**
  - **Add explanatory text to indicate what those paper clips signify!**
  - **Save file after validating that the file still conforms to PDF/A-3b**
  - **Digitally sign and lock file**
  - **Again, validate PDF/A-3b conformance**









Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

Consumer Utility Statements

- All this can readily be done with *current* tools and standards
- There is no good reason why these techniques cannot and should not be utilized by utilities and other vendors *now*
  - More secure
  - Lower cost of customer support (especially if receipts attached)
- In the future, such files may concurrently be tagged as complying with *both* PDF/A-4 (for archiving) and PDF/X-6 (for printing)

Adobe

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

STEM Document Equations

… It started off as an innocuous e-mail from an RIT graphic arts professor who passed along a request to me from an RIT computer science professor querying about heuristics for recognizing and extracting equations from an existing PDF file …

- **There is a demonstrated need to be able to extract equations (mathematical expressions) from PDF files for repurposing including indexing, experimentation, etc.**
- **Scraping is particularly difficult given the many methods of representing the "final form" of mathematical expressions**
  - **Text, vector, raster, or combinations of same**
  - **Symbols synthesized with multiple glyphs**
  - **Non-standard font encodings**
  - **Interesting spacing and scaling considerations**

Adobe

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

STEM Document Equations

- **Better to create PDF with data embedded representing these equations**
- **But what and how?**
  - **Attached file**
    - Each equation’s representation is relatively small
    - Potentially very many such objects in a document
  - **Metadata**
    - Requires PDF viewer that is equation metadata-aware

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

STEM Document Equations

- Competing formats
  - MathML - another W3C recommendation
  - TeX - a favorite of mathematicians using TeX for layout
  - Proprietary formats (Word, MathType, etc.)
  - Multiple files or a file with multiple formats for each equation
- Presentation versus Context;
  MathML provides both forms
- Can we get a consensus for what is really needed and actually useful?

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

STEM Document Equations

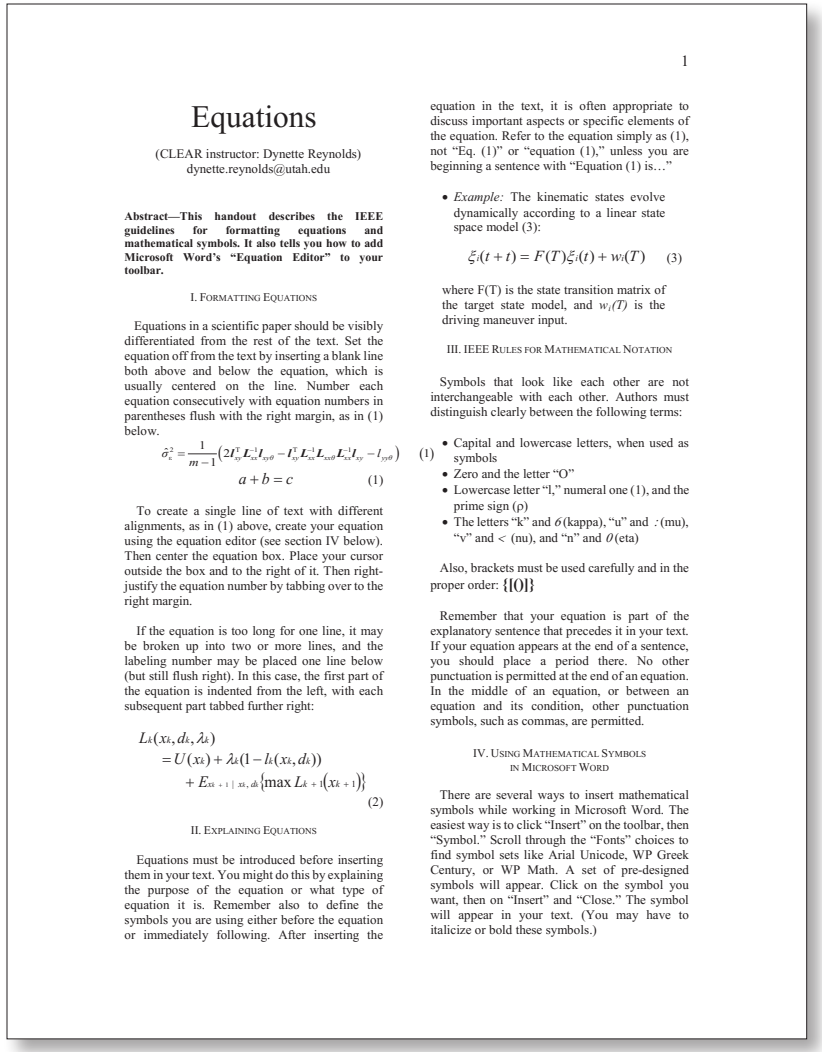
- **Example using similar techniques utilized with PG&E statement**
  - **Microsoft Word document with MathType-style equations**
  - **Created PDF file via Acrobat PDFMaker (including source file)**
  - **Embedded the files representing multiple formats available for equations (copied from MathType editor) via *attachment annotations* at appropriate locations within the document's PDF file**
    - TeX
    - MathML 2.0
      (both "m namespace" and "namespace attr" variants)
  - **Manually attached .EPS files saved from MathType for equations 1 and 5**
  - **Created a PDF/A-3b file via Acrobat Pro DC Preflight profile *Convert to PDF/A-3b (without fallback conversion)***

Adobe

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

STEM Document Equations

www.pdfa.org



1

Equations

(CLEAR instructor: Dynette Reynolds)
dynette.reynolds@utah.edu

Abstract—This handout describes the IEEE guidelines for formatting equations and mathematical symbols. It also tells you how to add Microsoft Word's "Equation Editor" to your toolbar.

I. Formatting Equations

Equations in a scientific paper should be visibly differentiated from the rest of the text. Set the equation off from the text by inserting a blank line both above and below the equation, which is usually centered on the line. Number each equation consecutively with equation numbers in parentheses flush with the right margin, as in (1) below.

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{m-1}\left(2\boldsymbol{I}_{xy}^{\mathrm{T}}\boldsymbol{L}_{xx}^{-1}\boldsymbol{I}_{xy\theta} - \boldsymbol{I}_{xy}^{\mathrm{T}}\boldsymbol{L}_{xx}^{-1}\boldsymbol{L}_{xx\theta}\boldsymbol{L}_{xx}^{-1}\boldsymbol{I}_{xy} - I_{yy\theta}\right) \quad (1)$$

$$a + b = c \quad (1)$$

To create a single line of text with different alignments, as in (1) above, create your equation using the equation editor (see section IV below). Then center the equation box. Place your cursor outside the box and to the right of it. Then right-justify the equation number by tabbing over to the right margin.

If the equation is too long for one line, it may be broken up into two or more lines, and the labeling number may be placed one line below (but still flush right). In this case, the first part of the equation is indented from the left, with each subsequent part tabbed further right:

$$\begin{aligned} L_k(x_k, d_k, \lambda_k) &= U(x_k) + \lambda_k(1 - I_k(x_k, d_k)) \\ &\quad + E_{x_{k+1}\,|\,x_k, d_k}\{\max L_{k+1}(x_{k+1})\} \end{aligned} \quad (2)$$

II. Explaining Equations

Equations must be introduced before inserting them in your text. You might do this by explaining the purpose of the equation or what type of equation it is. Remember also to define the symbols you are using either before the equation or immediately following. After inserting the equation in the text, it is often appropriate to discuss important aspects or specific elements of the equation. Refer to the equation simply as (1), not "Eq. (1)" or "equation (1)," unless you are beginning a sentence with "Equation (1) is…"

- *Example:* The kinematic states evolve dynamically according to a linear state space model (3):

$$\xi_i(t+t) = F(T)\xi_i(t) + w_i(T) \quad (3)$$

where F(T) is the state transition matrix of the target state model, and $w_i(T)$ is the driving maneuver input.

III. IEEE Rules for Mathematical Notation

Symbols that look like each other are not interchangeable with each other. Authors must distinguish clearly between the following terms:

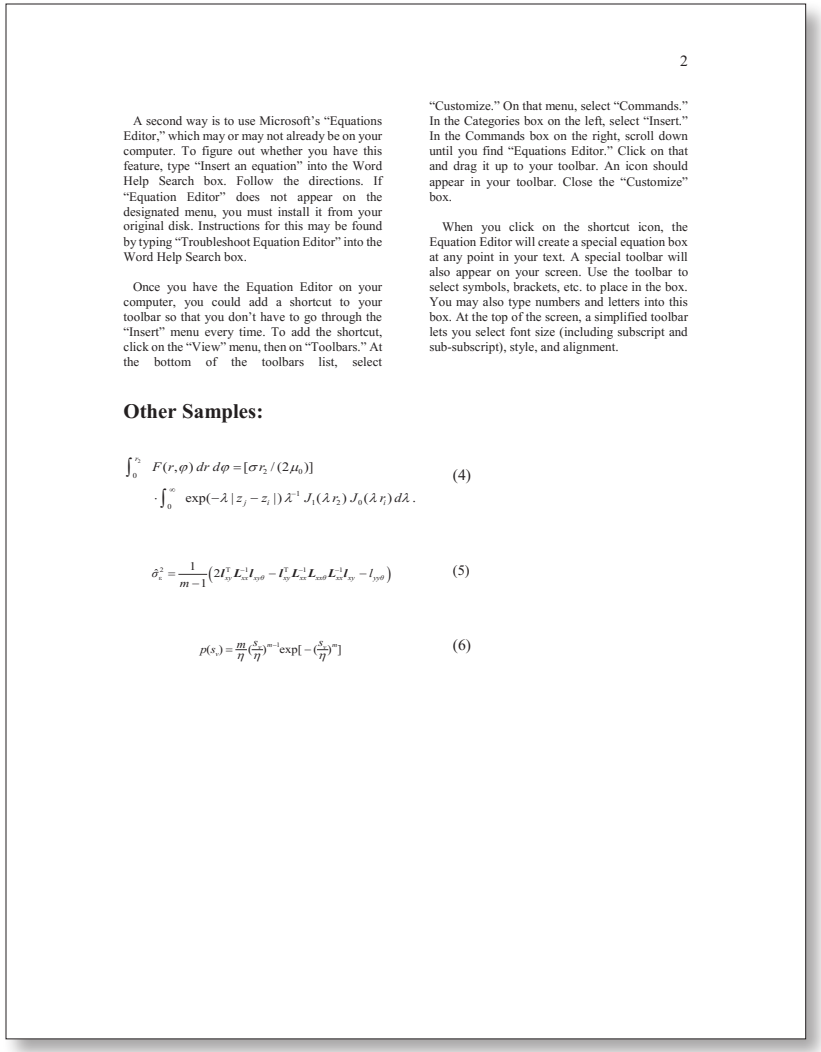
- Capital and lowercase letters, when used as symbols
- Zero and the letter "O"
- Lowercase letter "l," numeral one (1), and the prime sign (ρ)
- The letters "k" and *ó* (kappa), "u" and *:* (mu), "v" and < (nu), and "n" and *0* (eta)

Also, brackets must be used carefully and in the proper order: **{[()]}**

Remember that your equation is part of the explanatory sentence that precedes it in your text. If your equation appears at the end of a sentence, you should place a period there. No other punctuation is permitted at the end of an equation. In the middle of an equation, or between an equation and its condition, other punctuation symbols, such as commas, are permitted.

IV. Using Mathematical Symbols in Microsoft Word

There are several ways to insert mathematical symbols while working in Microsoft Word. The easiest way is to click "Insert" on the toolbar, then "Symbol." Scroll through the "Fonts" choices to find symbol sets like Arial Unicode, WP Greek Century, or WP Math. A set of pre-designed symbols will appear. Click on the symbol you want, then on "Insert" and "Close." The symbol will appear in your text. (You may have to italicize or bold these symbols.)



2

A second way is to use Microsoft's "Equations Editor," which may or may not already be on your computer. To figure out whether you have this feature, type "Insert an equation" into the Word Help Search box. Follow the directions. If "Equation Editor" does not appear on the designated menu, you must install it from your original disk. Instructions for this may be found by typing "Troubleshoot Equation Editor" into the Word Help Search box.

Once you have the Equation Editor on your computer, you could add a shortcut to your toolbar so that you don't have to go through the "Insert" menu every time. To add the shortcut, click on the "View" menu, then on "Toolbars." At the bottom of the toolbars list, select "Customize." On that menu, select "Commands." In the Categories box on the left, select "Insert." In the Commands box on the right, scroll down until you find "Equations Editor." Click on that and drag it up to your toolbar. An icon should appear in your toolbar. Close the "Customize" box.

When you click on the shortcut icon, the Equation Editor will create a special equation box at any point in your text. A special toolbar will also appear on your screen. Use the toolbar to select symbols, brackets, etc. to place in the box. You may also type numbers and letters into this box. At the top of the screen, a simplified toolbar lets you select font size (including subscript and sub-subscript), style, and alignment.

Other Samples:

$$\int_0^{r_2} F(r,\varphi)\,dr\,d\varphi = [\sigma r_2/(2\mu_0)] \cdot \int_0^{\infty} \exp(-\lambda\,|z_j - z_i|)\,\lambda^{-1} J_1(\lambda r_2)\,J_0(\lambda r_i)\,d\lambda\,. \quad (4)$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{m-1}\left(2\boldsymbol{I}_{xy}^{\mathrm{T}}\boldsymbol{L}_{xx}^{-1}\boldsymbol{I}_{xy\theta} - \boldsymbol{I}_{xy}^{\mathrm{T}}\boldsymbol{L}_{xx}^{-1}\boldsymbol{L}_{xx\theta}\boldsymbol{L}_{xx}^{-1}\boldsymbol{I}_{xy} - I_{yy\theta}\right) \quad (5)$$

$$p(s_i) = \frac{m}{\eta}\left(\frac{s}{\eta}\right)^{m-1}\exp\left[-\left(\frac{s}{\eta}\right)^m\right] \quad (6)$$

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

Summary

- PDF offers excellent opportunities and features not only to contain, represent, and deliver final form visual content, but also underlying data in usable formats
- PDF open data solutions are available now!
- PDF is absolutely ***NOT*** where data goes to die!

Adobe

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated

Thank you! Any questions?

Adobe

e-mail: <isaacs@adobe.com>

Dov Isaacs,
Principal Scientist,
Adobe Systems Incorporated