Using Social Media Tools For Data Science and Journalism to Communicate the Benefits of Open Government Data

Presentation for the W3C eGov Meeting
Dr. Brand Niemann
Director and Senior Enterprise Architect – Data Scientist
Semantic Community
http://semanticommunity.info/
AOL Government Blogger
http://gov.aol.com/bloggers/brand-niemann/
August 6 2012
Introduction

• I am using social media tools for data science and journalism to communicate the benefits of open government data to the public and decision makers:
  – Social Media Tools
    • Doing Linked Open Data and Semantic Analytics
  – Data Science and Journalism
    • Coming together in the Press Office
  – Benefits of Open Government Data
    • Semantic Medline Example of Real Benefit to US Agencies
  – Public and Decision Makers
    • Educating Them
Some Questions and Answers

• Who does the best?: The Guardian (Simon Rogers).
• What tools do they use?: Drupal?, Google Spreadsheets, Free Visualizations Tools, etc.
• What results have they produced?: See next slides.
• How can one use/reuse those results?: See my work in the next slides.
• What does this mean for our upcoming AOL Government Conference? What I will say.
Simon Rogers

• Simon Rogers is editor of the guardian.co.uk/data (http://www.guardian.co.uk/datablog, http://www.twitter.com/GuardianData), an online data resource which publishes hundreds of raw datasets and encourages its users to visualise and analyse them - and probably the world’s most popular data journalism website. He is also a news editor on the Guardian, working with the graphics team to visualise and interpret huge datasets. He was closely involved in the Guardian's exercise to crowdsource 450,000 MP expenses records and the organisation's coverage of the Afghanistan and Iraq Wikileaks war logs. He was also a key part of the Reading the Riots team which investigated the causes of the 2011 England disturbances.

• Previously he was the launch editor of the Guardian's online news service and has edited the paper's science section. He has edited two Guardian books: How Slow Can You Waterski and The Hutton Inquiry and its impact. In 2010, Simon received a special commendation from the Royal Statistical Society in its awards for journalistic excellence. His Factfile UK series of supplements won a silver at the Malofiej 2011 infographics award and the Datablog won the Newspaper Awards prize for Best Use of New Media, 2011.

• In 2011, Simon was named Best UK Internet Journalist by the Oxford Internet Institute, Oxford University and won the inaugural XCity award from City University. The Datastore was honoured at:
  – Knight Batten awards for innovation in journalism, 2011
  – Technical innovation, Online Media Awards 2011 Best use of new media for Guardian Datablog,
  – Newspaper Awards 2011

• Author of Facts are Sacred: the Power of Data (on Kindle) http://amzn.to/t2dAnR

http://www.guardian.co.uk/profile/simonrogers
Simon Rogers on Data Journalism

• But the game-changer for data journalism happened in spring 2010, beginning with one spreadsheet: 92,201 rows of data, each one containing a detailed breakdown of a military event in Afghanistan. This was the WikiLeaks war logs. Part one, that is. There were to be two more episodes to follow: Iraq and the cables. The official term for the first two parts was SIGACTS: the US military Significant Actions Database.

– Simon Rogers, The Guardian

The WikiLeaks War Logs (The Guardian)

NOTE: This version of the Draft Data Journalism Handbook is done in MindTouch, the leading social knowledge base, in one page so it can be easily searched as Linked Open Data!
NOTE: I use social media tools to collect, analyze and communicate data science for data journalism using a 5 steps process:
1. Build a Knowledge Base in MindTouch
2. Build Indices of the Knowledgebase in Excel
3. Precondition the Structured Data
4. Import the Indices and Structured Data into Spotfire and Create Visualizations
5. Write the Data Story following the DELTA Model* as follows:
   1. D: Accessible, high-quality data
   2. E: An enterprise orientation
   3. L: Analytical leadership
   4. T: A long-term strategic target
   5. A: A cadre of analysts

* Strategic Use of Analytics in Government (Sirkka L. Jarvenpaa and Thomas H. Davenport)
DataBlog – Data Store

Get The Data

NOTE: This needs to be parsed and is not the complete data!

https://docs.google.com/spreadsheet/ccc?key=0AonYZs4MzlZbdFdJWGRKYNvhWIB4S25OVmZhN0Y3WHc&hl=en_US#gid=0
Data Summary


NOTE: This has to be screen-scraped and reformatted!
Step 1: Build a Knowledge Base in MindTouch
The Real “All Our Data Journalism in One Spreadsheet”

Step 2. Build Indices of the Knowledgebase in Excel
Step 3. Precondition the Structured Data

NOTE: These had to be edited to work in Spotfire

NOTE: This also include tabs for the Olympic Medal history data that was scraped and formatted from the PDF poster file!

http://semanticommunity.info/@api/deki/files/18739/Guardian.xlsx
Analytics of the Data Journalism Spreadsheet in Spotfire

Step 4. Import the Indices and Structured Data into Spotfire and Create Visualizations

NOTE: Filtered by Spreadsheet to find stories with the most Number of Comments and Retweets for further analytics.

http://semanticommunity.info/AOL_Government/The_Guardian_Data_Journalism#Spotfire_Dashboard
Analytics of the Olympic Gold History Data Spreadsheet in Spotfire

http://semanticommunity.info/AOL_Government/The_Guardian_Data_Journalism#Spotfire_Dashboard
Visualizing Data: Guardian Shows How With Olympic Medal History

Those in the data community looking for new ways to express complex concepts might find this "Atlas of Olympic Success", visualized for the Guardian newspaper, to be a great example.

Excerpt from the Data Journalism Handbook

• The data feature **Eurozone debt web** explores the tangled web of intra-country lending. It helps to explain a complicated issue in a visual way, using colour and proportional arrows combined with clear text. An important consideration is to encourage the user to explore the feature, or follow a narrative, and never feel overwhelmed by the numbers.
  
  — Bella Hurrell and Andrew Leimdorfer, BBC

• MY NOTE: This was not true in my experience so I re-did their work.

Digital Agenda for Europe

http://semanticommunity.info/AOL_Government/Digital_Agenda_for_Europe
http://gov.aol.com/2012/06/29/digital-agenda-for-europe-data-as-first-class-citizen/
Eurozone Crisis: Deficit, GDP, Unemployment, and Debt in Spotfire

http://semanticommunity.info/AOL_Government/Digital_Agenda_for_Europe#Spotfire_Dashboard_3
38 Degrees: An AOL Gov Series
Conference On Data Innovation

38 Degrees is the first annual government and industry conference from AOL Government. Just as AOL Government covers the latest insider analysis and perspective on innovation at work in government via the Internet, 38 Degrees brings those ideas to life Sept. 19, 2012, at the Ronald Reagan Building in Washington, D.C.

This year’s inaugural event will focus on the vast and relatively untapped gold mine of freely available government data, and how the public – developers, entrepreneurs, enthusiasts and the venture capital community – can tap into its hidden potential.

38 Degrees represents more than the geographic latitude

http://gov.aol.com/2012/07/02/38-degrees-an-aol-gov-series-conference/
Example of Eric Busboom, CEO, Clarinova

• **Request:** I'm working on building a public data warehouse. The Project's goal is to make a wide variety of public datasets, cleaned, modeled and linked, available to the public through SQL and OLAP interfaces. Here is an overview of the project: [http://civicknowledge.org](http://civicknowledge.org)
  
  – The project seems well aligned with much of the work you have done at the EPA and at Semantic Community. It would be really valuable if I could talk to you about the project and learn more from your experience.

• **Response:** Get your investors to help you provide data science and data scientists to grow their businesses. This is what Greylock Partners does! I can help you with the work.
  
  – This will serve to introduce my editor and organizer of the conference, Wyatt Kash. I think you should definitely attend and have an interesting story to tell.