Statipedia: a platform for collaboration across statistical agencies

Peter B. Meyer and James A. Buszuwski

U.S. Bureau of Labor Statistics[1]

Meyer.Peter@bls.gov and Buszuwski.James@bls.gov

January 2010

1. Introduction

The United States government has a decentralized statistical system in which ninety or more agencies produce official statistics for a wide variety of data users.[2] Coordinating these many data producers is an important ongoing activity calling for continual efforts to maintain communication. The statistical offices have communicated for a long time through personal meetings, conferences, publications in scientific journals, and telephone calls. More recently, they have adopted communication technologies such as email and web sites.

Some agencies of the federal government now work together online. For instance, the agencies of the US intelligence community now use blogs, wikis, and instant messaging on a common platform that is available to their staff. This set of tools makes up a common "collaboration services" platform, sometimes referred to by the name of its main wiki, Intellipedia. Along these lines, we propose developing a platform of tools and workspaces for the staff of the agencies of the statistical community. It would make wiki software, source code control software, search engines, statistical programming languages, and other tools readily available to the employees who conduct research or provide expert opinion.

These new tools offer some important benefits. First, these technologies have been found to be an effective means of sharing information as new products and ideas are being developed. Traditional journals are also an effective means of sharing information, but mostly after the development is completed. The interactive nature of online technologies allows individuals to participate more easily in a development process as it occurs. Individuals can apply their different capabilities to the parts of complex problems for which they have comparative advantages or have particular interests or skills.[3]

Second, these technologies preserve a historical record of development. This makes it easy for an individual or team to accumulate source materials and ideas, search them, experiment with different ways of organizing them, and to keep track of past work. Those records then provide materials for forensic, training, and historical purposes.

Third, cost savings are to be expected by sharing these systems across agencies, which reduces the setup and maintenance costs. The key new technologies are available as "open source". Open source software has lower acquisition costs than a purchase of proprietary software, although the agencies would still incur costs for installing, maintaining and supporting the software.

Based on these observations and our interviews and experiments we believe that the government's statistical system could benefit greatly by adopting these online tools. This platform which we call Statipedia would make collaboration easier and more efficient. New projects to address new opportunities can arise from small groups of self-identified founders or moderators of a flow of source materials. Because outside users can inspect and copy the source materials, they can improve them and perhaps join a group. On the world-wide Web, this process generates new and improved open source computer programs.

The many common issues faced by the different statistical agencies present opportunities for collaboration. This may include wikis dedicated to specific issues such as seasonal adjustment, sampling, imputation, non-response, and classification

---

[1] Views and findings in this research do not represent official views, findings, or policy of the U.S. Bureau of Labor Statistics. This work draws heavily from an internal agency report on open source practices. The authors got valuable advice from many, many people; see Appendix A for our acknowledgements to them.

[2] Based on the list at http://www.fedstats.gov/agencies/ visited July 29, 2009.

[3] Noveck (2009, especially pp. 36-43) describes how and why *collaborative* governance, which enables opportunities to participate, represents progress beyond the *deliberative* governance process of listening to various viewpoints.

systems for occupations, industries, and products. A Statipedia platform could provide a forum for discussing new initiatives such as measuring "green jobs," or analyzing systemic risks in the financial system. It also may include blogs by recognized experts in an area of statistics or economics, or, perhaps, by agency representatives who would care to share their visions for the future of the statistical system.

Government statistical staff are sometimes constrained from using a tool they want to use for doing their work unless it is specifically hosted by their own agency. A Statipedia platform could mitigate this constraint if it could securely provide many services to users. One possibility for doing this is to offer all the Statipedia services through the new GSA collaboration site http://apps.gov.

2. Collaborative tools across the U.S. intelligence community

The events of 9/11/2001, and the disputes about the existence of Iraqi weapons of mass destruction in 2003, caused considerable concern about the coordination of information among the 16 agencies comprising the federal government's intelligence community. This concern prompted Congressional investigations that called for better and more streamlined communication leading to more accurate information and timely analysis.

The problem of communication in an environment of secrecy had been addressed by these agencies before. A separate top-secret military network had been set up around 1990, and a prototype network called Intelink among the intelligence agencies had been set up in 1994 partly in response to delays in making intelligence information usable in the field during the first Gulf War.[4]

It was against this background that analysts from the intelligence agencies investigated the use of Web 2.0 information technology as a means of meeting Congress' mandates.[5] Their goal was to create an environment for open internal discussion that would avoid excessive hierarchical or political control over the facts and interpretations which could be discussed across the agencies. Their work has resulted in a set of collaboration services that is now referred to as Intellipedia.

Intellipedia refers to a set of wikis, but the wider collaborative services platform includes blogs, instant messaging, chat services, document and video storage, and a search engine. The various pieces of software are mostly open source, with significant additions made by contractors and programmers in the intelligence community. Access is restricted to selected staff at the 16 intelligence agencies and the list is expanding over time.

Most of these tools were written as open source software.[6] The intelligence agencies developed a certification process to ensure that new versions are safe to use on intelligence networks. The agencies also have their own extensions or code changes, but they keep these to a minimum to make it easier to adopt new releases.

Intellipedia runs on MediaWiki, the same software that supports Wikipedia, but is administered differently. In the Intellipedia wikis, all edits are attributable to the person who logged in to Intellipedia. Wikipedia presents neutral, encyclopedic points of view, but in Intellipedia, an entry might make an argument or deliberately extreme interpretation. Users who think their views are underrepresented on a page can change the page or add a link to their own pages.

The experience of the intelligence agencies shows how government agencies can use open workspaces to work with confidential, collaborative, complex material.[7] Former Director of National Intelligence Michael McConnell reported that:

---

[4] Martin (1999) and interviews of Steven Schanzer and Franklin White by the authors and Jean Fox in July 2009. Schanzer was involved in setting up a military network for top secret information (JWICS) around 1990 and both worked to set up the intelink network for the intelligence agencies around 1993-4.

[5] Andrus (2005) was a key inspirational paper.

[6] An Oct, 2009 memo by David Wennergren, the acting Chief Information Officer of the Defense Department, nicely characterizes how and why open source software naturally meshes with the public service and government work. See attachment 2 at http://powdermonkey.blogs.com/files/2009oss.pdf.

[7] A complex or changing situation may call for many eyes and minds to interpret and analyze it effectively. Complexity and change also make it less realistic to specify in advance a fixed set of requirements on the system that addresses it. See also Brooks (1987) on computer projects and Alic (2007) on military projects. Our colleague Jason Ford has written about this in internal work at BLS, including Meyer et al (2008). Complexity and change are recurring attributes of both intelligence matters and statistical ones; so a knowledge-system for these fields must be flexible in ways that data-processing systems are

"Such tools enable experts from different disciplines to pool their knowledge, form virtual teams, and quickly make complete intelligence assessments."[8]

3. Statipedia platform tools

We propose a "Statipedia" for storing, sharing, and discussing government statistical methodology. The Intellipedia is a good model for this cross-agency cooperation. Both intelligence and statistics production are descriptive activities designed to observe, report and analyze evidence.

As we envision it, in its initial phase of development the Statipedia platform needs to include several tools and services:

1) Wiki software that supports scientific work.
2) A source code version control program with minimal licensing impediments, and shared source code repositories across the government.
3) A search capability across wikis, blogs, and other content on the Statipedia platform.

The tools are intended for storing and discovering source materials and collaborating on government work. For government workers they would complement, not duplicate or replace, Wikipedia, StatLib and other public web sites.

Over time if there is enough user interest, the platform could include other services such as blogging software. This would encourage discussion of current topics or provide insight on thinking about statistical issues. The platform could also offer access to statistical software on its "cloud," which would save on installation and maintenance costs and risks. A facility for storing user provided videos would be a way to preserve lectures, training seminars, and panel discussions from conferences. Access to the platform would start with government staff and might over time be granted to international statistical agencies, academics, and consultants.

Tool: Wiki

A wiki is a collection of web pages which can be edited right from the browser that views them. This technology realizes the original vision for the World Wide Web of both writing and reading directly on the web.[9] There are more than 50 implementations of wiki software listed at http://wikimatrix.org and http://en.wikipedia.org/wiki/Comparison_of_wiki_software.

On the central platform it will be useful to have a unified encyclopedia of government statistics and wikis for various topics, projects, and offices. These can run on several different kinds of wiki software, and have different layouts and configurations. Part of the value of having a platform with a lot of capacity is that the wiki pages, blogs, discussion boards and so forth can hyperlink to one another.

We have looked at a number of wikis for government organizations, including the intelligence agencies, the EPA, BLS, DoD, OMB, Eurostat, and the OECD. In the scientific context we looked at openwetware.org, scholarpedia.org, and Wikipedia. Most of these run the same basic software which is called MediaWiki. Installations can be customized with 400 extensions (or "plug-ins") to this software. Valuable features in the scientific context include footnotes, flexible category systems, embedded R programs, and mathematical expressions formatted in the TeX language.[10]

---

not. When there are a multitude of experts and stakeholders, different kinds of analyses are needed simultaneously, which is a reason for openness.

[8] This quote comes from a statement for the record of the Senate Homeland Security and Governmental Affairs Committee, 10 September 2007, http://dni.gov/testimonies/20070910_testimony.pdf.

[9] Berners-Lee (1999) described his 1990 vision of a read-write web where people would routinely post and read pages. He was surprised when so few people posted pages.

[10] Tools for writing mathematical documents often adopt the standard TeX language for formatting mathematical expressions. By developing more content in this standard, we make it easier to transfer content to and from specialists in this area. The conferences and journal of the American Statistical Association anticipate receiving papers written in the TeX format. Wikis with mathematical expressions allow TeX input and this is important for the shared platform, and makes it possible to collaborate on drafts of scientific papers right on the wiki. Another standard for communicating mathematics is MathML, which is a low-level XML specification for describing mathematics (http://www.w3.org/Math/). It is practical as a

Tool: Search engine

A search engine built into a single wiki or blog generally cannot find content at other wikis or blogs. Therefore the platform needs a search engine, possibly several, to search across all platform content. Then a user can find project wikis referring to a specialized topic such as "imputation" or "seasonal adjustment" or "capital measurement" with the relevant communities of experts. One tricky aspect of its configuration is that the search engine should only report content that the user has permission to access.

Tool: Source code version control

Much of what statistical agencies do is write and run software. We think their staffs should have access to certain software development tools such as version control systems.

Source code control software keeps track of files for software development projects. The source control software makes previous versions of each file available, and keeps track of who checked them in and when and what comments the person left. It keeps track of which files are associated with others, and allows batch processes to build new editions of the software. It keeps track of who has "checked out" a file and prevents others from checking a new version in. The interface to the stored code library is usually through a web browser. Little or no special software is needed on the client computer, although it is convenient if the user's text editor knows how to get versions from and save them to the code repository. Such "hooks" connecting the editors and source control tools are common.

Source code control software is used in software development environments around the world, but is not available to federal staff in general. However, now many white collar workers are writing computer programs. We expect the needs and opportunities for source control software to grow. It will be helpful for a source control system to be available to the staff by default, without having to ask permission or install anything on the user's own computer. The authors of a computer program can decide whether to make the code available to others or not. Big payoffs can come from projects that snowball once multiple groups can work together.

It is possible to offer more than one such service. The best known one is called 'subversion.' Others that were recommended include 'git' and 'bazaar.' An important benefit to having these available across the federal government is that it would reduce the number of redundant procurements and costs. As with our other tools we do not think the use of a particular source code control installation should be mandated. Instead, it is better to allow the clients to choose what to use for their own purposes, although there are efficiencies from having just one such system.

As we understand it, GSA's apps.gov will be offering the 'subversion' program through collab.net's service soon. We understand that for a 25-person site, the cost will be $5000 per year. At forge.mil the military has a highly fault-tolerant and redundant implementation of this service and it may be possible in the long run to use that platform for statistical work.

More tools

Many more tools and workspaces will be useful, and whenever we discuss this platform we receive many suggestions. So for example, if there were a collaborative platform we would like it to offer:

- mapping software which could be integrated with government data
- statistical programming software
- a repository of documents, with version control, perhaps usable for official documents
- a discussion board, like Govloop.com or Max.OMB.gov
- A questions-and-answers site perhaps modeled on stackoverflow.com
- Videos with training materials and past lectures and seminars

4. Demonstration web site

---

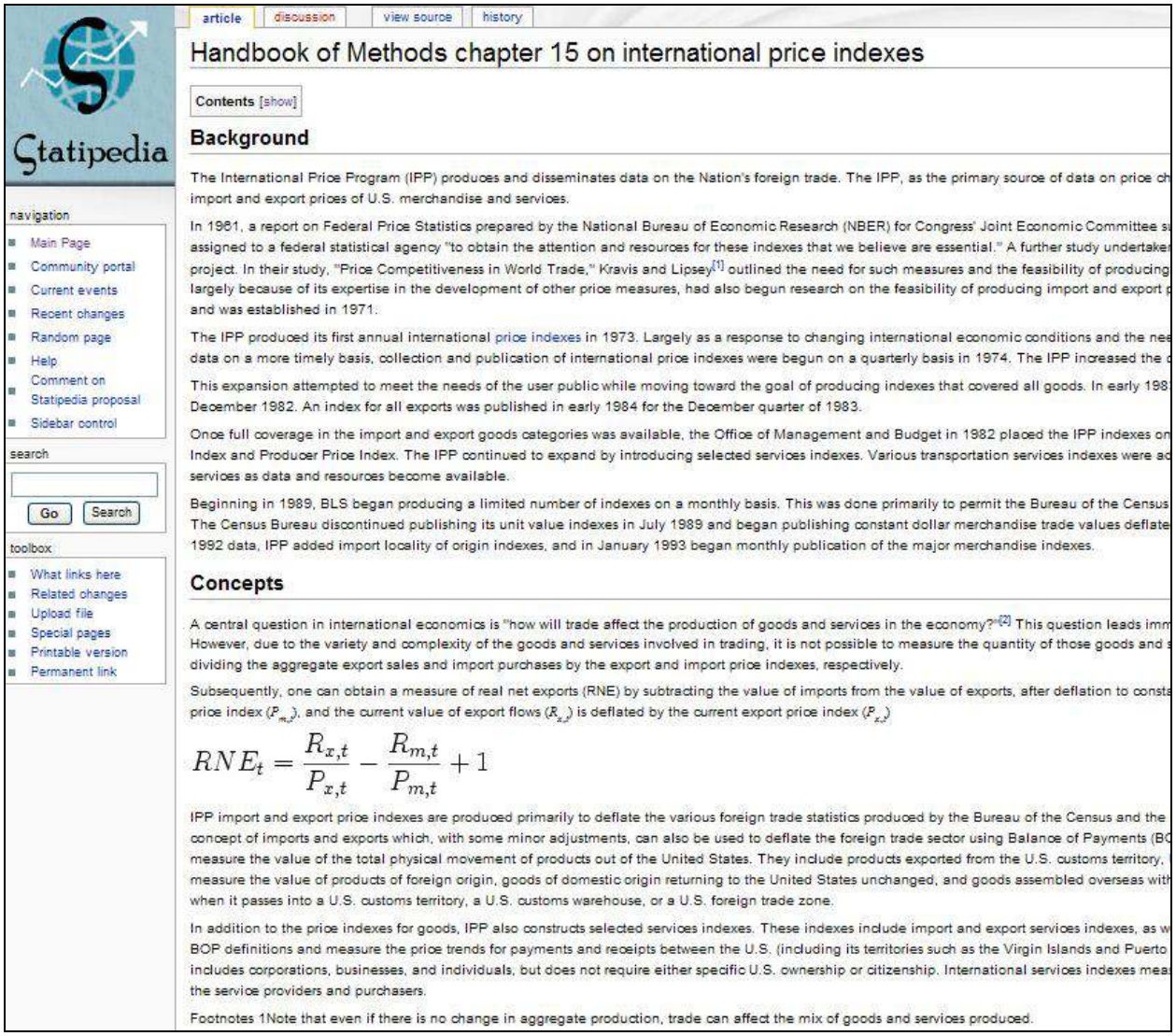
basis for machine to machine communication and can more accurately communicate the meaning of an expression to a machine, but is too verbose for users to comfortably edit directly. A search at the American Statistical Association site finds dozens of web pages referring to TeX but few referring to MathML.

We have made a demonstration site at https://statipedia.org to illustrate the usefulness of a platform like this. It has several elements for a demonstration:

Official documented methodology

We put a chapter of an agency's official handbook of statistical methods on this wiki. In a wiki format the text can become a more enriched living document, with a discussion page, newly inserted hyperlinks to definitions, and other ad hoc updates and added footnotes. The equations can be copied from this form into other documents as image or based on their source expression in the TeX formatting language.

Wiki pages with official content, such as this one, can be protected against edits by others.

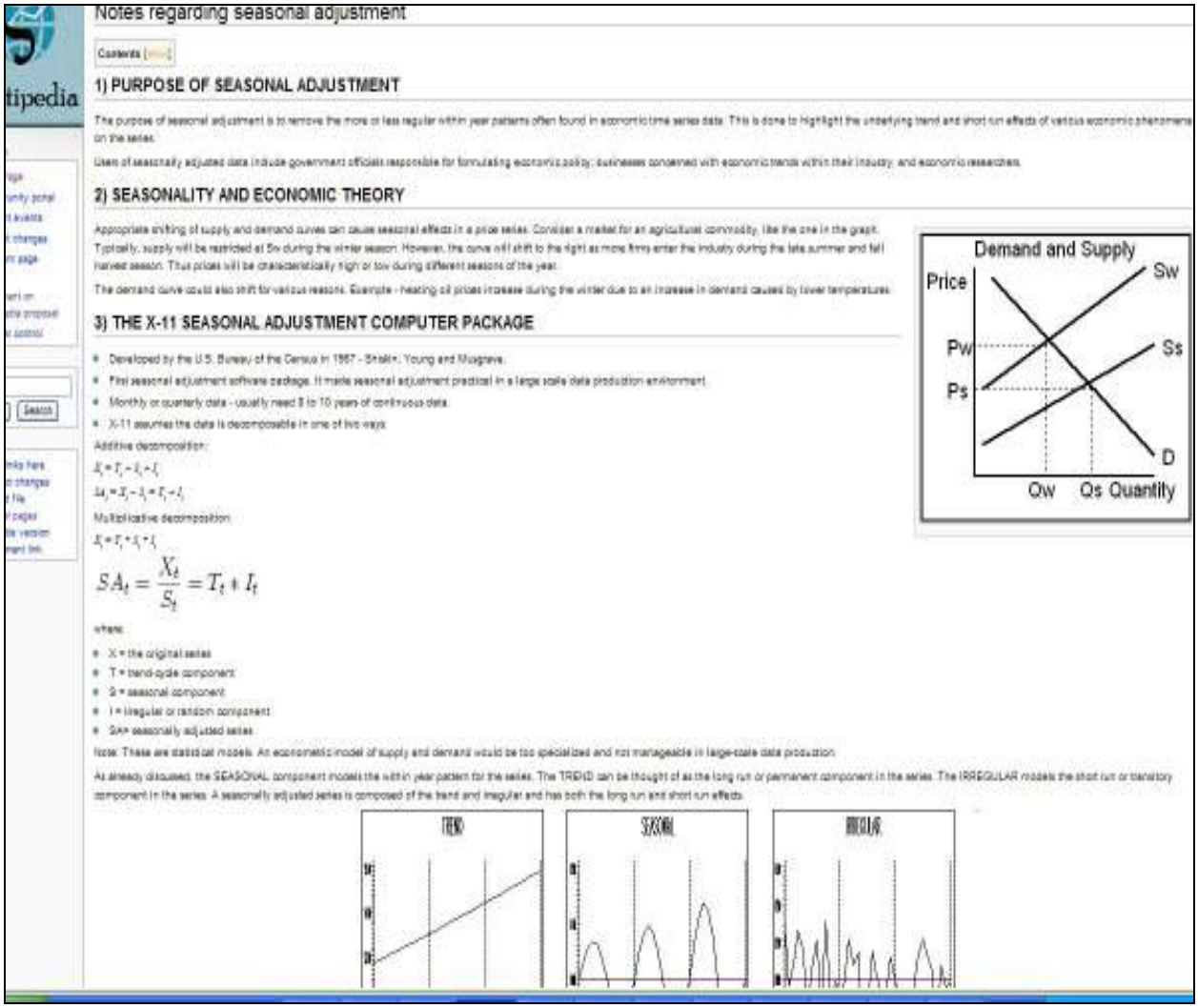


A page from an official publication copied onto a wiki

Training materials for seasonal adjustment

One of the authors teaches methods to do seasonal adjustment of time series. By putting training materials on the web, they can be reached by more people, both readers and editors. Specialized terms can be hyperlinked to definitions. Code examples can be shared by putting computer code on the web, and perhaps allowing it to execute there.

A person at another statistical office could straightforwardly benefit from the existence of the page. By searching the Statipedia wiki(s) for a term like “seasonal adjustment” they would find that such a page exists. The user could then copy and modify the text, equations, graphs, and arguments to work out a proposal for an index of crime statistics or seasonal adjustment of transportation statistics -- a different application from the original text. The second person might be in the same building, or across the country.

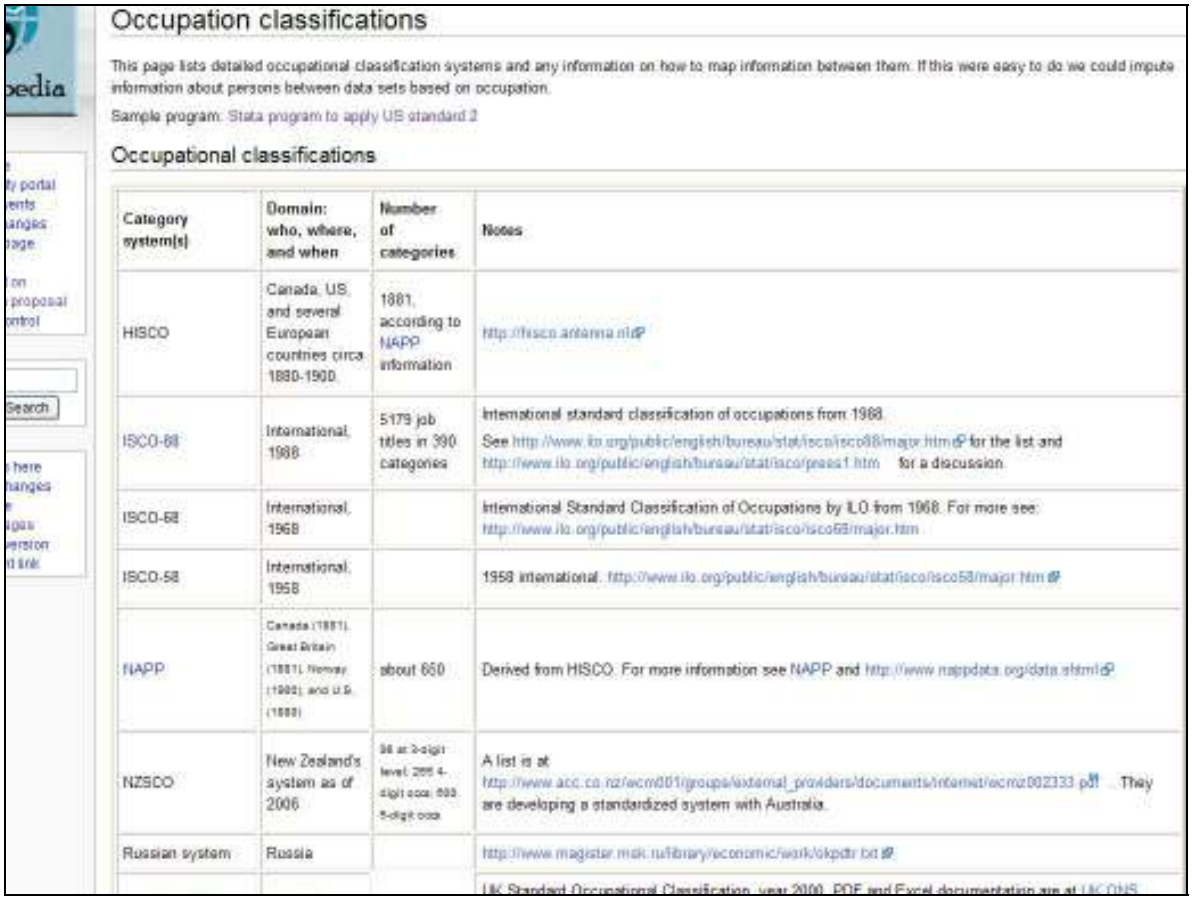


Training materials on a wiki

Information on past category systems

Many agencies have experts on the category systems used for official statistics or research. Official standards are developed in multi-year collaborations across government agencies and internationally. Historical and international category systems are developed by academics or in government. Occupations, industries, regions, and medical conditions have all had substantial histories of classification and revision over time.

In the case of occupations, some occupation category systems are designed to classify and link to education and training data, others to employment or compensation data, others to long term historical evidence, and others internationally. On the test site we list some historical standard occupation category systems. It is quite incomplete, but illustrates the range of past practices that an occupation expert might want to be able to rapidly refer to. Once it became a collaborative page on a government wiki, it could grow quickly with the wide information available to the users of the system.



Occupation classifications

This page lists detailed occupational classification systems and any information on how to map information between them. If this were easy to do we could impute information about persons between data sets based on occupation.

Sample program: Stata program to apply US standard 2

Occupational classifications

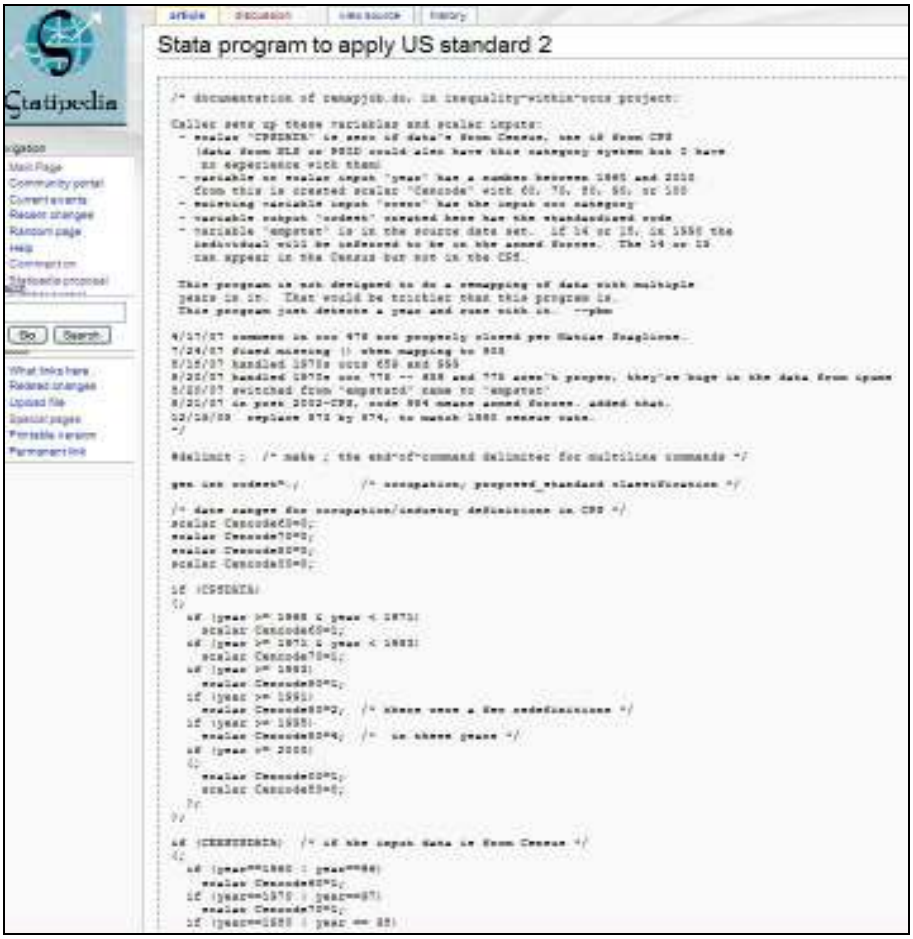
Category system(s)	Domain: who, where, and when	Number of categories	Notes
HISCO	Canada, US, and several European countries circa 1880-1900	1881, according to NAPP information	http://hisco.antenna.nl/
ISCO-88	International, 1988	5179 job titles in 390 categories	International standard classification of occupations from 1988. See http://www.ilo.org/public/english/bureau/stat/isco/isco88/major.htm for the list and http://www.ilo.org/public/english/bureau/stat/isco/praes1.htm for a discussion.
ISCO-68	International, 1968		International Standard Classification of Occupations by ILO from 1968. For more see: http://www.ilo.org/public/english/bureau/stat/isco/isco68/major.htm
ISCO-58	International, 1958		1958 international. http://www.ilo.org/public/english/bureau/stat/isco/isco58/major.htm
NAPP	Canada (1881), Great Britain (1881), Norway (1900), and U.S. (1880)	about 650	Derived from HISCO. For more information see NAPP and http://www.nappdata.org/data.shtml
NZSCO	New Zealand's system as of 2006	96 at 2-digit level, 295 4-digit occs, 693 5-digit occs	A list is at http://www.acc.co.nz/ecm001/groups/external_providers/documents/internet/ecmz002333.pdf. They are developing a standardized system with Australia.
Russian system	Russia		http://www.magistr.msk.ru/library/economic/work/okpdtr.txt
			UK Standard Occupational Classification, year 2000, PDF and Excel documentation are at UK ONS

Table of occupation classifications for reference and collaboration on a wiki

Sharing computer programs

In the next picture we illustrate how one might share computer programs. A wiki page about a subject can link directly to another wiki page with relevant computer code, or more advanced cases to a source control program's web page. Users could learn from the source code and consider reusing it for their own purposes. They can compare current versions to past versions.

The Defense Department has a shared source code control site at forge.mil for its own purposes and a statistical agency analogue would work similarly but with different content. For example, a number of government offices make price index calculations and do seasonal adjustment. For another example, the main Danish statistical office has a program to monitor blogs that refer to their agency in real time (Statistics Denmark, 2008). Some U.S. statistical offices might want to imitate this program if they could get it.



Computer source code for remapping occupations between category systems on a wiki

Definitions

Often definitions of terms vary slightly from agency to agency. The site has some examples which make this clear. Basic terms like "wages" or "the employed" are measured in different ways in different places. A definition which shows the range of variation and choice would give readers a practical sense of the measurement issues, and could help them choose which measure to use for a particular purpose.

Future examples: time-critical topics

In a time of crisis such as the shifting financial crises of 2008, shared tools and workspaces enable colleagues across government to work together rapidly and refer quickly to one another's experiences without a detailed authorization plan. By

increasing the communication among staff members and across offices, shared tools and workspaces may make catastrophic outcomes less likely (Edmondson, 2007).
One can therefore imagine a wiki for policymakers that would be useful for dealing with issues and events that unfold in real time such as those in the intelligence field. For example, recently the topic of "systemic risk" in a financial system has been very important. Some issues that evolved were: 1) how does one measure, model, or simulate systemic risk? 2) should this be done bank by bank or insurance by insurance company? 3) what is their capacity to withstand particular events or declines in asset prices? 4) how can we use what statistical agencies know and produce to address this issue in real time? 5) how does the Black-Scholes options theory relate to the empirical findings on systemic risk?

Time may be short in a financial panic, pandemic, weather event, natural disaster, mass migration, or terrorist attack. In such a crisis, the impediments associated with peer review in publication or communication through *channels* can delay useful information from "bubbling up" and being put to use in time. Approvals and publication are slow processes which formalize a text when its raw form may be more timely or relevant. By contrast, information and sophisticated opinion can develop openly and rapidly on a web *platform*. A common platform across organizations makes it practical to bring diverse cognitive resources to a problem area quickly[11] and it enables new ways of communicating in a crisis. For a platform to be available to policy makers when they need it, it must be set up and its format and content familiar to specialists, analysts, and programmers in advance.

Security features

When using URLs with the https:// prefix, transmissions between the web server and the browser are encrypted, so that devices at intermediate locations on the Internet cannot easily read the content. Banks use this on their sites. Our Statipedia.org site uses a commercial implementation of this protocol which costs less than $150 to set up and less than $20 a year to maintain. The wiki site is password-protected and the entry of the password can occur after the encrypted interaction begins, so that the password cannot easily be detected by a bad actor monitoring the transmissions. Because there is no secret content, visitors can view the demo site without logging in, but it would be easy to change this.

5. Policies, norms, and effects of Statipedia

Policies

As a general principle, policies governing Statipedia must serve cross-agency sharing of information and focus on large scale issues and concerns. The policies should also support opportunities for specialists to work together on a given topic. Below is a tentative list of policies for Statipedia that can start the discussion. As Statipedia evolves we can revisit these policies as the need arises.

1. Statipedia's design will focus on subject matter common to the statistical agencies. Wikis and blogs should usually be topical. Some examples might be a wiki for imputation, or one for seasonal adjustment. However, if a user feels that an alternative approach might be more natural when starting a new wiki or blog then we would also encourage them to proceed. For instance, a new wiki or blog might be motivated as representing the point of view of an organizational unit or that of an interagency project. But we would encourage users to be careful to avoid replicating an existing organizational structure with its local language and stovepipes that might exclude interacting with outsiders. (Don Burke expresses this principle by saying that a posting should be written for the widest possible audience.)
2. Our focus is to serve members of the scientific community. It is essential that the tools Statipedia offers to this group should be well adapted for the purpose of developing a knowledge base. In particular, users need some capability to improve the system, as scientists, technologists and analysts do in other environments. This includes extending the software, and substituting or upgrading modules of the system overall. (Noveck, 2009 includes such skilled functional participation in her concept of collaborative governance, which is not limited to the sharing of viewpoints.)

[11] Employment and compensation information may be useful in diagnosing a dot-com boom and crash; a real estate boom and crash; or a financial markets boom and crash. Thus Department of Labor expertise may be relevant to financial crisis managers. Joy's law applies: No matter which organization is responsible for the problem, most of the smartest people about that problem work somewhere else.

3. Statipedia would complement and aid the work of staff in the statistical offices. Staff should be encouraged to provide content and comments. But, we also believe participation should not be required as it may well meet resistance, friction, and delay. It is wiser to provide a basic system, then with experience evaluate its usefulness and the need to expand or constrain it. Statipedia will not take official decision making power away from the proper authorities as now constituted, nor directly replace existing computer systems.
4. If possible, we would like Statipedia's platform to be administered by a focused federal technology center, sponsored centrally perhaps by OMB. This will lift the burden of system management (e.g. upgrades, downtime, security issues, etc.) from the statistical agencies and reduce costs overall. GSA's new http://apps.gov platform offers an appropriate design, but specific suppliers for the various services are needed.
5. The tools provided will need to scale up to wide and expanding use, and may also need to be adapted by the specialists who use them in government. Proprietary licensing restrictions can prevent this or make it costly. Therefore, we recommend the platform be designed to use open source software generally although it can offer suitably licensed proprietary software when it is cost-effective.
6. Legal environment:
    a) The Clinger-Cohen Act gives agency CIOs (Chief Information Officers) authority over the use of computer systems, so an agency's staff may use it only with CIO permission. Partly because they create risk for these CIOs, cross-agency platforms are not yet common, but if use of the platform becomes standard practice, agency heads can help make approval routine.
    b) Some of the platform's contents would be subject to Freedom of Information Act (FOIA) requests, although these will probably be rare. Such requests can be satisfied by giving the requester copies of the relevant content, not access to the live platform itself.
    c) There is a risk of violating fair-use rules if copyrighted resources were to be on a common platform. We recommend that instead platform users hyperlink to or provide summaries of copyrighted source material. Federal-government-written content is generally not copyrighted.[12] For further information we can consult the cross-agency collaboration at cendi.gov.

Norms to encourage

Different communication media call for different interaction assumptions. On an interactive platform we cannot expect a level of rigor associated with scientific journals. Instead, our goal is to maintain a sense of informal professionalism among users that will facilitate sharing information and discussion of ideas. Here are a few guidelines for users which were written for an internal blogging facility at BLS:

1. Focus on work product and add value. Provide worthwhile information and perspective. Criticism is welcome in blogs and comments but should be offered in a constructive spirit, focused on helping to understand difficult issues and to identify areas of possible improvement. Providing concrete suggestions for improvement is encouraged.

2. Try to provide documentation for source material or evidence. Hyperlinks to data sources or documents are encouraged. Simple footnotes with citations help readers in their research. Providing relevant equations or source code on a wiki page can also be useful. Please take care to respect copyright and fair use laws.

3. Please identify yourself. If you are speaking as an expert, give some background information for people who do not know you. Links to user pages on a wiki help to establish credentials.

4. Blogging or using a wiki page can be a good method for seeking information on certain unusual or undocumented issues. However, you should try to make it a standard operating procedure to search source materials on existing wikis or blogs, and to link to them when appropriate.

[12] See the law at http://www.law.cornell.edu/uscode/17/105.html and a discussion at http://en.wikipedia.org/wiki/Copyright_status_of_work_by_the_U.S._government (referenced July 29, 2009).

5. Keep track of comments on your postings and try to respond when appropriate. Be open to the points of view of others. When quoting, highlighting, or criticizing another person, consider alerting them so that he or she has the opportunity to respond.

6. Know and follow your agency's official policies and rules for using information technology. Nothing in Statipedia replaces or changes them.

7. Keep it short if you can. Avoid requiring the reader to see every word to get the key points. An outline or table of contents can help. Also consider using the journalism convention of putting the most important things at the beginning. A lively wiki is made of *fragments* not completed *documents*. If you are posting a long paper and you would like comments, or if you are simply providing information to your fellow bloggers, then provide an overview paragraph preceding the paper.

8. Respect your audience and always use a high degree of professional decorum. Show proper consideration for others' privacy and for topics that may be considered objectionable or inflammatory. Phrase comments gently and clearly; comments posted to a blog can take on a life of their own.

Expected and desired effects on the statistical system

We hope for two main effects of the Statipedia platform on the statistical agencies. First, it makes useful tools available to more government staff. Secondly, by design it facilitates and enhances collaboration among the staff of the statistical system.[13]

"Transparency" is a desirable goal for government, stated by the current administration (OMB, 2009), and is a reason to create the Statipedia platform. But transparency of every detail of science in government is not compatible with normal white collar work. For example, if every email were public, it could put a chill on normal professional interactions and relationship-building. This also holds for statistical methodology as a work product. We do not want a platform that announces every minor error made in a government statistical office to the world. White collar workers do not wish to work under such circumstances, and there is little to be gained for the public by imposing a bright light on every failure.[14] Therefore, our hope is to improve "translucency" through and across the statistical agencies.

In the longer run, the existence of a common platform should encourage the conception and development of large-scale low-cost capabilities. For example, knowing Statipedia is available a staff member might think beyond his or her immediate purpose or agency when writing a program. Knowing that a new work can be made widely available to interested people, authors and developers can develop a good reputation by providing sound products. This provides the kind of incentive individuals need to participate on Statipedia and, in turn, provides benefits to other agencies who can then forego development costs to achieve similar results.

A centralized "cloud" service provider has lower installation and upgrade costs than the aggregated costs of many decentralized providers. For one thing, Statipedia systems could displace, over time, some existing systems which are more expensive per capita. For another, it encourages emergent standardization and collaboration in ways we cannot perfectly foresee today. For example, think of the cost of installing the software implementing the free statistical language R on many government computers. Simply making R available online ("in the cloud") would reduce that cost substantially.

---

[13] The National Academy of Public Administration report on collaboration (2009, pp. iv-v) describes well the effect the new forms of information technology can have on government. They enable a transition from rigid hierarchies to "collaborative communities of practitioners;" they enable governments to "gather ideas from a diverse variety of sources and filter the inputs based on value rather than origin;" and, they enable "empirical, data-driven decision-making."

[14] A valuable example of a transparency problem comes from a large project at Hewlett-Packard in 2001-2002 to create "corporate open source" software, whose code and project information would be visible more widely across the organization. For background, see Dinkelacker, Garg, Miller, and Nelson (2002) and Melian, Ammirati, Garg, and Sevón (2002). Some developers objected or resisted the changes, for several reasons including the sense that it invaded their privacy, e.g. by making incomplete work visible, by exposing errors in their English or programming, exposing their opinions unexpectedly, or exposing them to immediate direct competition from outsourced contract workers. This problem was called the "fishbowl effect." (Melian, 2007, e.g. pages 88, 194, and 201).

The Statipedia platform will enable and support scientific reproducibility of results (Stodden 2009). For example, an agency could provide statistical evidence from its studies and post the detailed tools and support materials. This would enable other agencies to learn from them, and perhaps motivate additional studies which would feed back benefits to the first agency.

Dissemination processes, such as listservs, generate flows of long documents by email, they are not efficient, and many of us are overwhelmed by them. A more desirable approach is to provide accessibility to information and to enhance their discoverability with appropriate tools such as search engines. The newer approach is to make documents available on networks and to provide a means to search them or to provide hyperlinks to them. We hope Statipedia will help to reduce the flow of email especially those with attachments.

Professional networking could also be improved by a platform like Statipedia by simply providing links to people who are oriented towards a research agenda. Networking could make it easy to find advice within the federal government or someone who might answer a quick question about an issue.

The knowledge-management opportunities that Statipedia offers are apparent in the context of an aging government work force. As workers with many years of service retire they take with them the best practices and rich contextual knowledge that may not be documented elsewhere. Statipedia could become a repository for this knowledge and, with an appealing wiki platform, new people in an organization could easily access this information and build on the work of their predecessors.

Long run prospects

In the long run we hope to work with many interested federal agencies on the Statipedia platform. We would also like to include some non-government participants who work closely with government economists. This includes university academics, consulting firms and trade associations. This implies that the platform would have to allow access from authorized non-government computers.

Another possible use for the Statipedia platform would be to make it an official repository for research and related documents. At the discretion of the appropriate agency, finished documents could be put on a site that is visible to the citizenry at large.

Other uses include hosting the work of statistical or computing user groups and providing data processing tools.

6. Getting Statipedia started

We believe the Statipedia platform should be sponsored by an organization with authority and expertise in science, policy, data collection, and data processing. The most logical is the Office of Management and Budget which has an Office of Statistical Policy. Other possible sponsors include the Bureau of Labor Statistics, the Federal Reserve Board, the new White House Chief Technology Officer's office, the Department of Commerce, or the Chief Information Officers Council. A formal Statistical Community of Practice is under discussion on OMB's discussion board, and this is a natural home. The OECD is beginning a wiki for the measurement of progress and this could offer similar services.

From interviews and discussions about projects of this kind,[15] we learned that projects to build platforms are more likely to launch if they are authorized by a high ranking official. The likelihood of conflict can be kept low by sticking with descriptive scientific subjects and not policy ones. The project should start small but with a scalable plan and technology, and it should offer new services rather than compete directly with existing systems.

Agreements across government agencies to work together are usually formalized in Memoranda of Understanding (MOUs). Such an agreement created CENDI.gov, a collaboration across federal departments with scientific and technical information.[16] There have been initiatives to do something similar for statistical agencies. We are seeking out interested parties and support for such an agreement and we are open to suggestions and advice.

---

[15] We have learned from conversations with two of the founders of intelink (Steve Schanzer and Frank White), several of the founders and developers of the Intellipedia (including Don Burke, Sean Dennehy, and Chris Musilek), and some of the managers and editors of the wiki publishing system at Eurostat, including Ulrich Wieland and Marc Debusschere.

[16] http://cendi.gov/about/history.html

7. Conclusion

We see much value in a Statipedia platform. With an open modular architecture, it can adapt over time in response to opportunities to add or upgrade its elements. We can move quickly to offer these services if we use the available open common standards most often used on the Web, in commercial environments, and in universities. By connecting the agencies better we can do better work, learn efficiently from the practices in other agencies, and save money on our existing work.

We can expect systems and interaction of this kind to raise the quality of government statistics, to reduce the costs of a given level of quality, and to raise the morale of the government staff doing this work. Previous internet systems such as Ethernet and the Web have had this effect. One of the rules of efficiency in this context is that it is helpful to adopt the external technical standards for the internal networks. This reduces the costs of interchange of data, skills, and information.

Appendix A. Acknowledgements

We got a lot of advice, expertise, and interaction from many colleagues, though they are not at all responsible for our phrasing or conclusions. We especially thank our colleagues on the Open Source Practices Team at BLS (Jason Ford, Jean Fox, Daniel Murphy, Curtis Reid, Daryl Slusher, Mark Thomas, and Elliot Williams) and BLS's Innovation Board. We benefited from advice from other colleagues at BLS including Dan Gillman, Steve Ferg, Richard Wallick, Stuart Scott, Richard Tiller, and Carl Barsky. We learned and were inspired by talking to experts from the U.S. intelligence community of agencies, including Don Burke, Sean Dennehy, Steven Schanzer, Frank White, Matthew Burton, and Chris Musilek. We learned much from colleagues at other agencies: Galen Pierce-Gardner (DOL/ETA), Mike Pulsifer (DOL/OPA), Andrew Felton (FDIC), Sarah Tricha (State Dept), Michael Messner (EPA), and Brian Monsell (Census).

We especially thank the organizers and participants at the conference on "Innovative Approaches to Turn Statistics into Knowledge" sponsored by the OECD, World Bank, and U.S. Census Bureau. Lynda Hawe told us about the OECD's wiki for progress. Ulrich Wieland and Marc Debusschere told us about Eurostat's wiki for publishing. James Forrester of the U.K. Cabinet Office advised us on the collaboration tools across the British government.

References

Alic, John A. 2007. *Trillions for Military Technology: How the Pentagon Innovates and Why It Costs So Much.* New York: Palgrave Macmillan.

Andrus, D.C. (2005). The wiki and the blog: Toward a complex adaptive intelligence community. *Studies in Intelligence*, 49(3), Available at SSRN: http://ssrn.com/abstract=755904.

Berners-Lee, Tim, with Mark Fischetti. 1999. *Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor*.

Dinkelacker, J., Garg, P.K., Miller, R., and Nelson, D. 2002. Progressive open source. *Proceedings of the International Conference on Software Engineering*, 177-184. http://portal.acm.org/citation.cfm?id=581363.

Edmondson, A.C. (2007). Mapping the Failure Landscape. *Presentation at Creativity and Entrepreneurship conference*, Harvard Business School, December 7, 2007.

Martin, Frederick Thomas. 1999. *Top Secret Intranet*. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Melian, Catharina 2007. *Progressive Open Source: the construction of a development project at Hewlett-Packard*. Stockholm School of Economics PhD dissertation. Economic Research Institute.

Melian, C., Ammirati, C.B., Garg, P., Sevón, G. 2002. Building networks of Software Communities in a Large Corporation. Hewlett Packard Technical Report. Presented at the *American Academy of Management Conference*.

Meyer, Peter B., James A. Buszuwski, Jean Fox, Daniel Murphy, Curtis Reid, Daryl Slusher, Mark Thomas, and Elliot Williams. 2008. "Open Source Practices Team Report". Bureau of Labor Statistics internal report.

Meyer, Peter B., and James A. Buszuwski, 2009. "Statipedia; a platform for collaboration across statistical agencies". Presentation at OECD, World Bank, and Census Bureau conference *Innovative Approaches to Turning Statistics into Knowledge*. http://www.oecd.org/dataoecd/37/35/42597793.pdf?contentId=42597794

Noveck, Beth Simone. 2009. *Wiki Government: How technology can make government better, democracy stronger, and citizens more powerful*. Washington, DC: Brookings Institution Press.

National Academy of Public Administration, Collaboration Project Advisory Panel (Greg Lashutka, P.K. Agarwal, William Eggers, Mark Forman, John Kamensky, Anne Laurent, and Franklin S. Reeder). "Enabling Collaboration: Three Priorities for the New Administration." January, 2009. Online: http://www.scribd.com/doc/11644716/Enabling-Collaboration-Three-Priorities-for-New-Administration.

Office of Management and Budget. OMB Open Government Directive. Memorandum M-10-06. December 8, 2009. http://www.whitehouse.gov/omb/assets/memoranda_2010/m10-06.pdf

Statistics Denmark, prepared by Rune Stefansson. 2008. Monitoring and reacting in the blogosphere and other online media. UNECE Work Session on Statistical Dissemination and Communication, 13-15 May 2008. http://www.unece.org/stats/documents/2008/05/dissemination/wp.2.e.pdf.

Stodden, Victoria. 2009. http://www.stanford.edu/~vcs/talks/VictoriaStodden-UOI-20min.pdf