First Monday, Volume 16, Number 2 - 7 February 2011

first monday
PEER-REVIEWED JOURNAL ON THE INTERNET

Open data: Empowering the empowered or effective data use for everyone?

by Michael Gurstein

Abstract

This paper takes a supportive but critical look at "open data" from the perspective of its possible impact on the poor and marginalized and concludes that there may be cause for concern in the absence of specific measures being taken to ensure that there are supports for ensuring a wide basis of opportunity for "effective data use". The paper concludes by providing a seven element model for how effective data use can be achieved.

Contents

The open data movement

The open data movement [1] in the area of access to public (and other) information is a relatively new but very significant, and potentially powerful, emerging force. It has now been widely endorsed by, among others, Tim Berners–Lee [2], often referred to as the father of the World Wide Web. The overall intention is to make local, regional and national data (and particularly publicly acquired data) available in a form that allows for direct manipulation using software tools as for example, for the purposes of cross tabulation, visualization, mapping and so on.

The underlying idea is that public (and other) data, whether collected directly as part of a census collection or indirectly as a secondary output of other activities (crime or accident statistics, for example) should be available in electronic form and accessible via the Web. There are significant initiatives in this area underway in the U.S. [3], the U.K. [4] and Canada [5] among many other jurisdictions and as part of a wide variety of not–for–profit initiatives as well.

This drive towards increased public transparency and allowing for enhanced data–enriched citizen/public engagement in policy and other analysis and assessment is certainly a very positive outcome of public computing and online tools for data management and manipulation. However, as with the earlier discussion concerning the "digital divide" there would, in this context, appear to be some confusion between movements to enhance citizen "access" to data and the related issues concerning enhancing citizen "use" of this data as part, for example, of interventions concerning public policies and programs.

Open data access vs. open data (effective) use

In an earlier paper dealing with the digital divide [6] discussion, I suggested the use of the concept of "effective use" to distinguish between the *opportunity* for digitally enabled activity presented by information and communications technology (ICT) access, from the actual *realization* of those opportunities in the form of "effective use". At that time I introduced a set of layers of requirements, which can be understood as "pre–conditions" for the realization of "effective use" of digital "access".

Efforts to extend access to "data" will perhaps inevitably create a "data divide" parallel to the oft–discussed "digital divide" between those who have access to data which could have significance in their daily lives and those who don't. Associated with this will, one can assume, be many of the same background conditions which have been identified as likely reasons for the digital divide — that is, differences in income, education, literacy and so on. However, just as with the "digital divide", these divisions don't simply stop or be resolved with the provision of digital (or data) "access". What is necessary as well is that those for whom *access* is being provided are in a position to actually make *use* of the now available access (to the Internet or to data) in ways that are meaningful and beneficial for them.

The question then becomes, who is in a position to make "effective use" of this newly available data?

The suggestion implicit in most of the discussions on "open data" (and explicit in Berners–Lee's previously quoted talk) is that "everyone" has the potential to make use of the data. However, as we know from experience elsewhere, not "everyone" has access to the digital infrastructure, to the hardware or software, or to the financial or educational resources/skills which would allow for the effective use of data or any other digital resource. Thus, rather than the entire range of potential users being able to translate their access into meaningful applications and uses, the lack of these foundational requirements means that the exciting new outcomes available from open data are available only to those who are already reasonably well provided for technologically and with other resources.

The example that Berners–Lee quotes concerning the role of the data mashup in the Zanesville lawsuit [7] is an interesting case in point. In this instance, the direct creators of the mashup were the Cedar Grove Institute [8], a public interest consulting firm specializing in GIS applications and employing several leading Ph.D. GIS specialists and with a University of North Carolina MBA as the CEO. The lawyer who argued the case, and presumably who so effectively deployed the mashup, is a Harvard law school graduate.

Of course, there is nothing wrong with this, nor with the outcome of their intervention and their use of open data — in fact, as with Berners–Lee, I think this is an exemplary case of the positive benefits for people that can come from open data.

However, this is a very long way from what folks like Berners–Lee seem to be asserting which is that "open data" empowers everyone. In fact, the example indicates precisely the opposite, that is, that "open data" empowers those with access to the basic infrastructure and the background knowledge and skills to make use of the data for specific ends.

These above mentioned resources are more likely to be found among those who already overall have access to and the resources for making effective use of digitally available information. This would then suggest that a primary impact of "open data" may be to further empower and enrich the already empowered and the well provided for. On the other hand, those most in need of the benefits of such new developments may find themselves out of luck (unless of course, they have means or the luck to find benefactors such as the Cedar Grove Institute or Harvard Law School graduates willing to work *pro bono* or on a contingency basis).

Raising critical issues concerning open data

A very interesting and well–documented example of this empowering of the empowered can be found in the work of Solly Benjamin and his colleagues looking at the impact of the digitization of land records in Bangalore. Their findings were that newly available access to land ownership and title information in Bangalore was primarily being put to use by middle and upper income people and by corporations to gain ownership of land from the marginalized and the poor. The newly digitized and openly accessible data allowed the well–to–do to take the information provided and use that as the basis for instructions to land surveyors and lawyers and others to challenge titles, exploit gaps in titles, take advantage of mistakes in documentation, identify opportunities and targets for bribery, among others. They were able to directly translate their enhanced access to information along with their already available access to capital and professional skills into unequal contests around land titles, court actions, and offers of purchase for self–benefit and to further marginalize those already marginalized [9].

Similarly, in response to my earlier blog on this matter Michael Spencer wrote the following to me in a private e–mail that "(You don't) have to go to Bangalore. The very same thing is happening in Nova Scotia where, over the last 20 years, there's been a concerted push to get all land titles, deeds, surveys and other data into a publicly available GIS. Because much (most?) of the rural land in NS is still unsurveyed" … people are rooting through various data being put online — nineteenth century deeds, ancient maps and so on … "then pay surveyors and lawyers to, in one way or another, take land away from owners. A neighbor, 80 yrs and illiterate, lost some 50 acres from the family homestead … All the old–timers in the community know the land belonged to him and to his daddy before him" but he lost it anyway. "(this) wouldn't have been economic (or even possible) … before the digitized GIS system went public."

Certainly the newly digitized information was "accessible" to all on an equal basis but the availability of resources to translate that "access" into a beneficial "effective use" was directly proportional to the already existing resources available to those to whom the access was being provided [10].

Benjamin's meticulously documented paper describing the process of digitization of land records and the aftermath in Bangalore, shows how the digitization and related digital access to land title had the direct effect of shifting power and wealth to those with the financial resources and skills to use this information in self–interested ways [11]. This is not to suggest that processes of computerization inevitably lead to such outcomes but rather to say that in the absence of efforts to equalize the playing field with respect to enabling opportunities for the use of newly available data, the end result may be increased social divides rather than reduced ones particularly with respect to the already poor and marginalized.

This is not to argue against "open data" which in fact is a very significant advance and support to broad–based democratic action and empowerment. Rather it is to argue that in the absence of specific efforts to ensure the widest possible availability of the prerequisites for "effective use" the outcome of "open data" may be quite the opposite to that which is anticipated (and presumably desired) by its strongest proponents.

An effective use approach to open data

An "effective use" approach to open data would thus be one that ensured that opportunities and resources for translating this open data into useful outcomes would be available (and adapted) for the widest possible range of users. Thus, to ensure the effective use of open data a range of considerations needs to be included in the open data process and as elements in the open data movement. This would include such factors as the cost and availability of Internet access, the language in which the data is presented, the technical or professional requirements for interpreting and making use of the data, and the availability of training in data use and visualization, among others.

An interesting example of how open data, with appropriate attention being given to some of these pre–conditions, in fact can provide a basis for effective use can be seen in how the UCLA Centre for Health Policy Research's California Health Interview Survey (CHIS) has been put to use by community advocates in Solano County. The CHPR conducts a bi–annual California Health Interview Survey [12] in conjunction with the California Department of Health "to provide a snapshot of the health and healthcare of Californians".

The survey is used by a range of political authorities. Free and widely accessible training is available on how to use the information "to develop appropriate and targeted policy responses" and overall "to learn how to use and apply the data to improve health and health care". That is, the information is not only made accessible, but attention is paid and resources are provided to ensure that the data is usable by those who might make effective use of it. In this instance, the Solana County community advocates were trained so as to be able to take the data provided by the CHIS, and plot incidences of asthma by local electoral district. They were then able to create a map showing an extremely high frequency of asthma among residents in a particular local area. The community advocates successfully argued against developing another truck stop along I–80 in the county based on CHIS 2001 data estimates that showed Solano County to have the state's highest rate of asthma symptom prevalence overall and one of the highest rates for children [13].

While in many respects this example parallels the earlier one from North Carolina the difference here is that the skills required for doing the analysis of the online data were provided through training to the local community who were then able to mount a local campaign to achieve the desired end. The key difference here was the attention that was paid by the provider of the information, the CHPR to ensuring that the data could be effectively used without the need for highly skilled (and expensive) professional intermediaries. This involved the development of end–user oriented training programs.

In this instance it should be noted that Internet access, bandwidth, the language of the data among other factors were not an issue. However, in other circumstances such as, for example, among indigenous peoples, non–English speakers, the very poor, and those living in areas with poor connectivity, these issues will be inhibitors of use of open data. A responsible intervener would be concerned to ensure that these issues were attended to as part of an open data program.

Additionally, the difficulties and types of interventions required to ensure that effective use can be made of information by the intended clients can be found in a very interesting report from Shelter in the U.K. [14]. This report documents a very useful approach to providing some of the tools needed for effective use of online information by those to whom that information is being directed and who would necessarily be those who could make the most active and effective use of that information — information on housing for the homeless being made available for use by the homeless themselves.

For many, if not most of those currently advocating for (and benefiting from) open data, access to the resources required for effective use is not generally a problem. Many of those advocating for open data are themselves professionals in the use of data for a variety of research and policy intervention purposes. What this means is that they have access to the bandwidth, the computers, the software, and the organizational structures required for effective use of the data or they have the access to the financial or other resources sufficient to allow them to obtain the supports for effective use of the data. In fact, the most likely immediate beneficiaries of open data are those with the most resources to make effective use of the data — the private sector who have the means and the interest in directly translating available data into new commercial products, or services or marketing strategies.

Open data supply side and demand side

This paper deals primarily with what might be termed the "demand" or "user" side of open data — what are the impacts and requirements for potential users in order that "open data" be truly "open" and "usable" by all. Tim Davies, in a comment on the original blogpost [15], discussed the "supply" side of open data; that is, what is required from the providers of open data so that it can be openly accessed by potential users. What Davies is referring to here, and in his Master's thesis [16] on the subject, and which are further reinforced by other comments on the blog, are those data related background conditions and data design decisions — formatting, (geo and other) tagging, selectivity, even initial research collection design — which directly impact on the nature and usability of the data. As well, these would to a degree, selectively impact on particular users or particular uses, having the effect of either enabling or disabling particular end uses or end users. These directly influence how the data which is being made available (open to the user) is structured, configured, and otherwise pre–processed prior to it being provided to the end user community.

A second commentator on the original blogpost, Zainab Bawa, makes a somewhat similar argument but relates the issues of the supply side of data more directly to the broader social and cultural context from which the data has been gathered (processed). Bawa further links this into how the now, decontextualized data is recontextualized in a new form both of which processes (decontextualizing and recontextualizing) have significant impacts on the semantic content of the data and thus on how the data gains meaning for the end user. One example of this would be how crime statistics gain a great deal of their meaning (and thus the effect of their use) from the geographical divisions through the data is formatted and made available (*i.e.*, where geographical boundaries are built into the data descriptions for example).

This leads to a third and perhaps even more important point raised by a number of commentators. In looking at "open data" it is necessary to include a three–step process: "access", "interpretation" and "use". In my blogpost I only referred to the "access" and "use" elements and made some non–articulated assumptions about matters of interpretation/meaning (or sense) making.

The point that is made here is that the process of interpreting or understanding “open data” is a separate process from making (effective) use of the data and that any critical analysis of “open data” use has to include how and under what conditions the data that is being made available is contextualized and given meaning. Thus, for example, the case drawn from Tim Berners–Lee’s presentation, the “interpretation” (sense making) of the data was the contribution of the consulting firm and was presumably based on their experience and expertise in ongoing work with geographically based information and advocacy.

In the Solana case, it was a specific feature of the UCLA Health survey to provide training in data interpretation. This training was then directly available for application by the Solana community to take their local (anecdotal) experience with high incidences of asthma and be able to interpret the data made available through the survey to give support to their advocacy. In the case of the land digitization in Bangalore (and the very interesting parallel example provided concerning digitization of land records in Nova Scotia, Canada) it was the expertise that was available to the wealthy landowners that enabled them to exploit the digitization process.

A model for effective data use

I have earlier [17] discussed a seven–layer model to achieve effective use of the Internet, beyond simple Internet access as a response to the digital divide. In this paper, I update this model in the context of responding to an anticipated “data divide”.

The following itemizes various elements that are required to be in place for end users to have the opportunity for “effective use” of open data. Some of these are more essential than others but to my mind some component of each needs to be in place or large numbers of those who might otherwise make use of open data to improve their lives and particularly the poor and marginalized will be excluded from making “effective use” of open data.

These include:

1. *Internet* — having an available telecommunications/Internet access service infrastructure sufficient to support making the data available to all users. Issues here would include:

   a. the affordability of Internet access — a major issue for many particularly in the developing world.
   b. the availability of sufficient bandwidth for the range of uses to which the data might be effectively put, for example, whether the data access has been designed on the basis that for example, broadband is necessary for the use of the data being made available.
   c. the accessibility of the network, for example, where access to the network or to connectivity is restricted for political or other reasons.
   d. physical accessibility/usability of access sites as for example for the physically disabled.

2. *Computers and software* — having access to machines/computers/software to access and process the available data and machines that are sufficiently powerful to do various analyses; having sufficient time on the equipment to do the analyses (many people need to share computers); knowledge of how to operate the equipment sufficient to access and analyze the data and so on. Does the use of the data require more powerful (and expensive) computers or software than might be generally available, for example?

3. *Computer/software skills* — having sufficient knowledge/skill to use the software required for the analyses/making the mashups/doing crosstabs, etc. Techies know how to do visualization, university trained persons and professionals know how to use the analytical software but ordinary community people might not know how to do either and getting that expertise/support might be either difficult or expensive or both.

4. *Content and formatting* — having the data available in a format (language, coding for display, appropriate geo–coding) to allow for effective use at a variety of levels of linguistic and computer literacy. What are the language, computer literacy, data analytic literacy levels that are required for an effective use of the “open data”? Does the use of the data presume that it is being used by a professional and are there means through which those professionals might be available to those who can’t afford expensive fees?

5. *Interpretation/Sense making* — sufficient knowledge and skill to see what data uses make sense (and which don’t) and to add local value (interpretation and sense making); being able to identify the worthwhile information and to figure out how to put the data into the right format or context so that what might otherwise be numbers on a page becomes something that can change people’s lives.

6. *Advocacy* — having supportive individual or community resources sufficient for translating data into activities for local benefit. Availability of skills and local resources, community infrastructures, training, the means for advocacy and representation all are required to enable effective local interventions based on the open or other data.

7. *Governance* — the required financing, legal, regulatory or policy regime, required to enable the use to which the data would be put.

Applying the effective data use model

Looking closely at the above list and then cross–checking it with the cases discussed earlier, it is clear that say in the Zanesville case, quoted by Berners–Lee, all of the elements are in place but many of them — and particularly the data formatting and analysis, interpretation, and advocacy (#3, #4, #5, and #6) — are being provided by expert professionals. In the Solana case the UCLA Centre is providing a degree of support for local application (#4) and targeted training for community advocates (#5). The community very likely, with the support of state funding, to turn data into advocacy (#6). Solana is located in a wealthy and highly developed part of the world, ensuring that they have access to the required infrastructure and software supports (#1, #2 and #3) and to a legal/regulatory system that is open to this kind of data driven advocacy (#7).

The wealthy landowners in Bangalore are, as a matter of course, able to provide themselves with the basic technical infrastructure of Internet access, computers and software (#1, #2 and #3). The government of India, through its digitization program, is ensuring that element data (#4) is available in a useable format and that there is a supportive legal and regulatory system for enforcing the outcomes of decisions and actions based on this data (#4 and #7). Given their financial resources, the wealthy landowners are able to hire professionals for elements interpretation and (self–interested) “advocacy” (#5 and #6). In the case of Bangalore again, even if there are publicly accessible means for gaining Internet access and computer use (#1, #2 and #3) and even though the actions of the Government of India provide what they would consider a “level playing field” for elements #4 and #7, in the absence of financial resources to interpret the data and then develop advocacy actions based on the data (#5 and #6), the poor and marginalized would be unable to use their data access in any meaningful way.

Conclusion

What the above analysis suggests is that for “open data” to have a meaningful and supportive impact on the poor and marginalized, direct intervention is required to ensure that elements currently absent in the local technology and social ecosystem are in fact, made available.

In the absence of such interventions, as Tim O’Reilly so correctly observed in tweeting my blogpost, not only can open data not be used by the poor but in fact “open data” can be used “against the poor”!

About the author

Dr. Michael Gurstein is currently Executive Director of the Centre for Community Informatics Research, Development and Training (CCIRDT) in Vancouver, Canada. He is the Editor in Chief of the *Journal of Community Informatics* (http://ci-journal.net) and Foundation Chair of the Community Informatics Research Network. He has held research professorships at universities in Canada and the U.S. and is an Adjunct Professor at the Information School at the University of Toronto. He has consulted to the governments of Canada, Australia, New Zealand, Malaysia, South Africa, Nepal and Jordan; to the Ford Foundation, Hewlett Foundation, U.N. Development Program, and European Union; and to Nortel, Mitel, Bell Canada, and Intel, among others. He has been on the Board of the Global Telecentre Alliance, Telecommunities Canada, the Pacific Community Networking Association and the Vancouver Community Net and is a member of the High Level Panel of Advisors of the (U.N.) Global Alliance for ICT for Development. His most recent book is *What is community informatics (and why does it matter)?* (Milan, Italy:

Polimetrica, 2007).
E–mail: gurstein [at] gmail [dot] com; blog: gurstein.wordpress.com

Acknowledgments

This paper represents a "recycling" of two individual blogposts with the addition of considerable material. The initial blogpost, specifically on open data, *went viral* (going from an average of 80 hits for each new blogpost to roughly 4,000 hits in two days and a total of 6,000+ hits to date!), after being commented on by several leading open data advocates and then being re–blogged and tweeted very actively. But perhaps more usefully it also got caught up in a form of "crowdsourcing" and a number of extremely useful comments and suggestions were made both on the blog and by e–mail discussing the overall notions in the blogpost. These have significantly advanced my thinking in this area — and so, thank you to all who commented and my apologies for not referencing all of you individually. The second blogpost, outlining the "effective use" approach has now been downloaded some 500 times.

Notes

[1] Wikipedia, at http://en.wikipedia.org/wiki/Open_Data, accessed 8 January 2011.

[2] This is a widely circulated speech given by Tim Berners–Lee at TED, the very well known tech community thinkfest. The site describes the talk as follows: "At TED2009, Tim Berners–Lee called for 'raw data now' — for governments, scientists and institutions to make their data openly available on the Web. At TED University in 2010, he shows a few of the interesting results when the data gets linked up." At http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide.html, filmed February 2010, posted March 2010, accessed 8 January 2011.

[3] The Data.Gov Web site (http://www.data.gov/, accessed 8 January 2011) describes itself as follows: "Data.gov is leading the way in democratizing public sector data and driving innovation. The data is being surfaced from many locations making the Government data stores available to researchers to perform their own analysis."

[4] The Data.Gov.UK Web site (http://data.gov.uk/, accessed 8 January 2011) describes itself as: "Over 5,400 datasets to view: Inside government data: Who's who in government and where does the money go?"

[5] Open Parliament.ca (http://openparliament.ca/, accessed 8 January 2011) describes itself as: "Info on what your representatives are doing in Ottawa can be hard to find and use. We're trying to make it easy."

[6] Michael Gurstein, 2003. "Effective use: A community informatics strategy beyond the digital divide," *First Monday*, volume 8, number 12, at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1107/1027, accessed 21 January 2011.

[7] This is the blogpost with extensive links to local reporting and the court record on the case referred to extensively in Berners–Lee's talk at TED cited above — http://www.zimbio.com/The+Coal+Run+Discrimination+Lawsuit/articles/12/UPDATE+ON+THE+COAL+RUN+LAWSUIT, accessed 8 January 2011.

[8] This is the Web site for the consulting group that supported the community organization (Coal Run) referred to in Tim Berners–Lee's TED presentation — ://www.cedargroveinst.org/, accessed 8 January 2011.

[9] Solomon Benjamin, R. Bhuvaneswari, and P. Rajan, 2007. "Bhoomi: 'E–governance', or, an anti–politics machine necessary to globalize Bangalore?" *CASUM–m Working Paper*, at http://casumm.files.wordpress.com/2008/09/bhoomi-e-governance.pdf, accessed 8 January 2011.

[10] A.J. Liebling remarked that "Freedom of the press belongs to those who own one." It equally applies here as in the nineteenth century.

[11] Benjamin, *op.cit.*

[12] California Health Interview Survey (CHIS), at http://www.askchis.com/, accessed 8 January 2011.

[13] http://www.chis.ucla.edu/pdf/chis_making_impact.pdf, accessed 8 January 2011.

[14] See http://scotland.shelter.org.uk/about_us/how_we_make_a_difference/housing_aid_and_advice/specialist_services_in_scotland/assisted_access_pilot_project
accessed 21 January 2011. This project was cancelled by the U.K. government.

[15] http://gurstein.wordpress.com/2010/09/02/open-data-empowering-the-empowered-or-effective-data-use-for-everyone/, accessed 21 January 2011.

[16] T. Davies, 2010 "Open data, democracy and public sector reform: A look at open government data use from data.gov.uk" (unpublished Master's thesis, University of Oxford), at http://practicalparticipation.co.uk/odi/report/wp-content/uploads/2010/08/How-is-open-government-data-being-used-in-practice.pdf, accessed 8 January 2011.

[17] Gurstein, *op.cit.*

Editorial history

Received 2 December 2010; revised 10 January 2011; accepted 20 January 2011.

Open data: Empowering the empowered or effective data use for everyone?
by Michael Gurstein.
First Monday, Volume 16, Number 2 - 7 February 2011
http://journals.uic.edu/ojs/index.php/fm/rt/printerFriendly/3316/2764