Perspectives on Open Data: Issues and Opportunties

Donald Cowan, *Member, IEEE,* Paulo Alencar, *Member, IEEE,* and Fred McGarry,

Technical Report CS-2014-01

Abstract—Open data like big data has become an important new direction in information technology. Governments are releasing data so as to be more transparent and also claiming that open data has substantial economic value. However, there appear to be many issues about open data that are not being discussed. This paper uses several practical examples in an attempt to illustrate many of these issues and allied opportunities, and through them to suggest a partial research agenda around open data.

Index Terms—open movement, open data, open data use cases

1 Introduction

Open data is based on the concept that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. However, republishing does imply citing the original source not only to give credit but to ensure that the data has not been modified or results misrepresented. The Open Data Movement (ODM) is growing rapidly among all levels of government and non-government organizations (NGOs) in many of the world's economies. Reasons for this interest include:

- enhanced transparency and accountability of the governments and agencies that release data;
- efficiency and improvements in Public Service delivery;
- enhanced inspection and collection of data through increased citizen engagement; and
- creation of economic and social value.

Current evidence [1], [2] strongly supports that each of these objectives creates value by either saving money or unlocking new business and social opportunities. This statement appears to be particularly true when data and knowledge from all levels of government, NGOs and business are integrated in specific applications.

Federal, state or provincial, and municipal governments in many countries have jumped on the open data bandwagon and are publishing data very quickly. The Canadian federal government has published approximately 200,000 data sets [3] since the move toward its open data policy.

What data should be published apart from that which might impinge on national security and privacy? How are we going to find, access and maintain such an amount of data effectively once the floodgates truly open? Thus, the release of open data raises numerous issues that need to be addressed. Some key questions are:

- How do you find open data that has been released to public view?
- How can open data be prepared so that it can be easily accessed?
- How do you develop applications so that viewing, comparison and use of open data can be made accessible to the broader public?
- How do you identify opportunities to create value from open data?
- How do governmental and non-governmental organizations with limited resources decide what data to make open and how to maintain it, once it is public?
- Will new technological approaches to open data be required much like those being considered for big data?
- Can the definition of open data include data, that has potential security or privacy implications, yet can be shared by community of practice networks?

The intent of this paper is to discuss these issues and related opportunities with a view to creating a research agenda that should be pursued, as open data becomes more pervasive. What is needed is an infrastructure for open data and a set of protocols and processes that govern how open data can be used and maintained. The amount of open data comes

- *Donald Cowan and Paulo Alencar are with the David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.*
E-mail: dcowan/palencar@csg.uwaterloo.ca
- *Fred McGarry is with the Centre for Community Mapping, Waterloo, Ontario Canada N2L 2R5.*
E-mail: mcgarry@comap.ca

from many different sources and is almost limitless, and thus like big data, is a victim of our exploding ability to capture, store, and share data of all kinds. However we are limited by our ability to curate,[1] search, analyze and visualize. In other words, once we have truly unleashed the open data monster, how do we maintain, find and use it?

2 The Complexity of Open Data - Use Cases

The power of open data to make government officials more accountable and improve public services such as health care has been illustrated by examples both from North America and Europe [1], [2]. However these applications are only the tip of the open data iceberg. There are many examples of the use of open data that could make our society not only more accountable but also more safe and secure from phenomena such as climate change.

In this section we describe three environmental examples, one land development example, one from our indigenous society and one from social and community support. These examples are provided in an attempt to illustrate the breadth and complexity of possible applications for open data to support societal security, societal and government accountability, community development and business. These examples are not hypothetical; they are existing software systems and related problems that are evolving as more is learned about open data applications and the interaction with various users. Each system, with one exception, is under development or was developed by the Computer Systems Group at the University of Waterloo (UWCSG) in partnership with the Centre for Community Mapping (COMAP).

2.1 Watershed Modeling

To illustrate one aspect of the use of open data we provide an environmental use case based on Canadian data, but we believe that the concepts apply equally in any jurisdiction. The software to support this use case is being developed by UWCSG and COMAP in conjunction with Greenland International Consulting Engineers and is under development as the paper is being written.

The example relates to the application of advanced research in building software modeling tools to understand the behaviour of watersheds and water catchment areas using a number of open data sources and existing predictive modeling techniques. Such tools can then be used in a decision support role to help in predicting a catchment area's behaviour under different scenarios related to current land use, planned land development or extreme weather events.

Extreme weather happened in several locations in Canada, the United States and Europe in 2013 causing billions of dollars in infrastructure and property damage and loss of life. These modeling and decision support tools in conjunction with open data can be used in both the long and short-term to predict watershed behaviour and thus prevent or ameliorate damage from extreme weather events, other environmental impacts and hopefully minimize loss of life.

These models require access to large amounts of open data related to the watershed being modeled including: catchment and stream delineation, digital elevation, soil texture, water holding capacity, erosion potential and soil drainage, weather station locations, daily precipitation, min/max temperature records and land use. The open data sets just described are used both as input to models, and to calibrate the models to ensure that the output is credible.

Of course, this open data, although existing in many cases, is scattered among multiple jurisdictions. For example, open data for a watershed in any Canadian province could be held by Federal Departments, Provincial Ministries (departments), conservation authorities,[2] political regions such as municipalities, NGOs, universities or consultants. Because the data is not accessible from a single source it becomes extremely costly to assemble, thus preventing the wider use of predictive simulation tools. How do we find all this data and provide access to it and how do the suppliers of the open data manage to keep it current even when they are a small municipality with a limited budget?

What appears to be needed is an environmental data and software platform operating as a cloud that is accessible over a high speed network. This platform will not only contain data and software but could be connected to field personnel and sensor networks that can deliver data through both satellite and land-based communications in near real-time.

2.2 Water Course Modeling

Streams and rivers are an important part of our environmental infrastructure. It is necessary to understand and model their behaviour under many different conditions. For example, runoff from the land can change the temperature of the water or change its chemical composition. In the first case, the change in temperature could change the ambient fish population; in the second case chemicals like phosphorous, a component of fertilizer, can deplete oxygen in the water, by causing excessive algae growth, thereby destroying all fish habitat.

As in the previous section all the data does not come from one source. In the Canadian context, monitoring of water courses such as streams and rivers

[1] select, organize and look after items in a collection

[2] In Canada conservation authorities are responsible for managing and maintaining watersheds that are in one province.

for fish populations falls under the provincial governments. Measurement of phosphorous deposition is the responsibility of the local conservation authority and land use is usually under the jurisdiction of a municipalities although often governed by provincial/state or federal laws.

The Flowing Waters Information System (FWIS) is a "Collaborative Geomatics" service developed by UWCSG and COMAP in partnership with the Ontario (Canada) Ministry of Natural Resources. The purpose of FWIS is to provide municipal planners and resource agency staff with access to collective stream fisheries data in the Lake Ontario basin to protect sensitive fish habitat. FWIS provides reasonably up-to-date and comprehensive fish species habitat and stream flow information from more than 1500 stations monitored by a dozen collaborating resource agencies. Access to this data allows practitioners to test future development scenarios to predict impacts on fish species habitat and stream conditions for streams where monitoring data is unavailable.

Recently (2013), 196,046 fish distribution records from 4230 sites and studies collected by 30 partners within 32 Great Lakes watersheds over the past 42 years were uploaded to FWIS. FWIS offers an approach for identifying: where and what data has been collected, who collected the data, and which protocols were used. This information will facilitate better science development, state of resource reporting, monitoring and data sharing.

FWIS is only the start. A wealth of information on Great Lakes Basin streams has been collected by numerous government and non-government agencies on both the US and Canadian sides of the basin. Periodically, researchers have compiled data to develop a better understanding of the influence of tributaries on the Great Lakes fisheries. Recent initiatives on both sides of the basin show tremendous progress towards understanding the influence of landscape conditions on fisheries. Additionally, progress is being made in better standardizing the collection of field data.

However, such analyses require a large investment in data preparation that involves soliciting numerous agencies that collect stream data and requesting their latest sampling data. Inevitably, researchers must repeat the process of gathering data for each new research initiative at tremendous costs and time. In Ontario, a network approach of collecting data and sharing it through a free open database has demonstrated that there is significant interest in both standardizing data collection and sharing survey results.

While the protocols may be comparable, there are still two issues with data management for this information. First, while government agencies are making progress at providing standardized systems, the resulting data are not generally easily accessible by other agencies. Second a large number of stream surveyors still store information on individual personal computers, or in spreadsheet format. As a result, it is difficult to access available information on tributary fisheries production in Great Lakes basins.

The availability of a common information system accessible to all stream surveyors would:

- reduce costs,
- increase communication and data sharing,
- increase sample sizes and power of statistical tests on the effects of regulations,
- reduce response times for regulatory planning processes that can impede economic development, habitat improvements or other management techniques, and
- permit comparisons to be made across spatial and temporal scales.

Moreover, a common information management system accessible to all stream surveyors through the internet could provide a venue to facilitate development of a community of practice on Great Lakes stream fisheries management. The availability of a collective data set that can readily be compiled from a web-enabled information management system will support predictive modeling.

2.3 Invasive Tracking System

Climate change and globalization in tandem are causing species to migrate into locations where they have not appeared before. The Asian carp, purple loosestrife and giant hogweed are three examples of these so-called invasive species that have been imported through the effects of globalization. In contrast the Mountain Pine Beetle has moved into new habitat because of warmer winters caused by climate change. These various invasive species are not only a nuisance but in many cases can be dangerous to humans and damaging to the environment. For example, Asian carp can threaten native freshwater species and touching a giant hogweed can cause painful blisters or even blindness, if it comes in contact with the eyes.

The problem then becomes one of taking corrective action such as control or eradication if possible. Of course the first step is locating and identifying the invasives. COMAP and UWCSG worked with the Ontario Federation of Anglers and Hunters (OFAH) to build a system for identifying and locating specific instances of invasives.

The concept is to build a mobile app that contains a field guide to invasives similar to a bird-watchers guide and an ability to record a picture of the specific invasive and its location using a smartphone or a tablet. Once a person has taken the picture and recorded the location with the onboard GPS the information is uploaded to the person's workspace accessible over the web. The observation can then be augmented and the GPS position adjusted to reflect the real location before submitting the observation to a master database.

Tracking of invasive species is not possible with government representatives; there are just not enough personnel. Thus, a form of crowd-sourcing must be used where members of the public interested in tracking invasives can record the occurrence, augment it and then submit the instance for final online vetting of the observation by experts. In this case vetted crowd-sourced data becomes open data and part of the public record.

2.4 Oak Ridges Moraine

The Oak Ridges Moraine is an ecologically important geological landform in the plains of south-central Ontario, Canada. The moraine covers a geographic area of 1,900 square kilometres. The Oak Ridges Moraine's hydrological system is a major constituent of the Humber Watershed supplying water to significant parts of Toronto, Canada, so that any factors affecting the moraine may have an impact on a major population centre.

Development on the land occupied by the Moraine is governed by the the Oak Ridges Moraine Act. However, with 35 upper and lower-tier municipalities on the Moraine, it is difficult to ensure that each of these communities is in compliance.

COMAP with UWCSG are developing a Web-based system with the Save the Oak Ridges Moraine Coalition to provide citizens of the 35 communities on the Moraine with an ability to monitor development approvals in each community on the Moraine with a view to ensuring compliance with the Act.

Obviously these systems require access to open data such as zoning bylaws, amendments to those bylaws and the minutes of community council meetings to ensure that the Moraine Act is being followed. Bylaws are a particularly difficult form of open data as they tend to grow erratically as amendments are piled upon amendments for areas or specific properties.

COMAP with UWCSG has also developed a Web-based system for the Caring for the Moraine project of the Oak Ridges Moraine Foundation. The Caring for the Moraine project enables 30 conservation-minded organizations to offer land-owners technical advice and access to various resources to undertake projects on their properties that help protect and restore the Oak Ridges Moraine. Land owners are able to access this information and also determine the land use designation, conservation priorities and available conservation resources for their properties. The system also supports communication among the various concerned groups who will be able to publish and post their news, events, exemplar projects and photos securely to a temporally and spatially searchable collaborative map of the Oak Ridges Moraine.

2.5 Aboriginal Atlas

Many political jurisdictions recognize that indigenous peoples have rights over the land they have occupied for many decades if not centuries, their so-called traditional lands. In Canada these rights have been clarified through Canadian Supreme Court decisions in the form of an action called "duty to consult." Duty to consult means that any substantive change to an indigenous group's traditional lands is governed by negotiation [4], [5] between the indigenous group and the proponent(s) of change. These negotiations often result in settlements such as an impact and benefit agreement that can provide monetary and other forms of benefits to indigenous communities. For example, the agreement might include provisions for quotas for jobs for local indigenous people, purchases from native businesses or other local economic opportunities.

The Mississaugas of New Credit First Nation (MNCFN), a Canadian indigenous people, in conjunction with UWCSG, COMAP and Shared Value Solutions (SVS) are developing services that will enhance the capacity of the MNCFN community to:

- Build a spatial database of their cultural heritage and environment;
- Respond to, manage and benefit from the impacts of land infrastructure and resource development in their traditional territory; and
- Reveal their history and connection to the lands for the edification of the general public.

This system called Dreamcatcher will form the foundation of a service for indigenous communities for the benefit of the MNCFN and client communities. Dreamcatcher will have the following attributes:

1) A spatial and document database with secure map and mobile GPS content contribution services to enable the MNCFN and their researchers to capture, delimit spatially and substantiate sites and landscapes valued by MNCFN. (Including, but not limited to: MNCFN traditional territory, MNCFN land claims, reserve lands, sacred grounds, resources used for traditional purposes, current resource use information, archeological sites of significance to the MNCFN, cultural heritage archival and community narrative information, constraint mapping which buffers areas of identified significance and access to open data services for ecosystem data maintained by external organizations for environmental modeling research.)
2) A secure web service that proponents can access to register their contact information, enter details for their proponent proposals and upload digital versions of supporting documentation and spatial information (shapefiles) to a searchable spatial document management system.
3) A mapping facility with highly resolute imagery, selectable thematic layers, and facilities for the import of proponent spatial data; that enables the MNCFN and their researchers to comprehend implications of proposals in the

spatial context of valued sites and landscapes and ecosystem information. A spatial negotiation service that enables and captures consultations with shapefile revisions by the MNCFN and proponents.

4) Workflow services that order, track and schedule the management, communications and documentation for each proponent proposal.
5) An administration service that provides role-based access to data, workflow and mapping publication services and community forum services for internal discussions on proponent proposals and valued sites and landscape information.

Although the indigenous groups will make use of open data and present their results in a fairly open manner (see point 1 in the previous list) they will not completely reveal their data but put a buffer around it. The data may involve points or areas of significance that relate to traditional medicinal plants or sacred sites to name two examples. This information has to be protected in some manner so that areas are not desecrated or destroyed or otherwise harmed.

2.6 Community Services Project

The following paragraphs in this section present a real open data problem that exists in many cities. Although our research team (UWCSG and COMAP) investigated the problem in conjunction with social service agencies in a city, there was no funding available to design and implement the underlying information system.

The goal of this project was to create and implement a community planning process using an open and accessible web-based asset map that puts mapping tools into the hands of the groups such as social service NGOs that have the least resources in the community. What would be captured are the services and corresponding service areas within a substantial area of the city. Thus, overlaps and gaps in service could be detected and acted upon thereby making an attempt to rationalize them.

Hopefully this exercise would provide better services at lower cost. Such a project could include established social service agencies, grassroots groups, businesses, faith groups, residents and any other interested community groups that live or work in the area being modeled. Social service agencies that have secure access to statistical data for purposes of gap analysis would use open data contributed by all participants but share restricted data within their secure community of practice application.

2.7 Summary

Each example in this Section illustrates issues with providing open data. None are insurmountable but significant human and technical resources are needed to address them.

Watershed modeling requires data from multiple government, NGO and private sources, which need to be identified, while water course modeling has an assembly problem as well. Dealing with management and control of development in an area with restrictions is particularly challenging even though the zoning bylaws and council minutes are already open data. There needs to be substantial effort in making bylaws and minutes consumable, perhaps relating them to a map of the community might be a first step.

Crowd-sourcing of data can also be an issue as described in Section 2.3. When can crowd-sourced data become valuable and in the official open record?

In the case of the indigenous peoples there is a distinction between truly open data and data, which can only be partially revealed for reasons of security. Data that has been collected might refer to a feature that would be attractive to tourists or might refer to sacred sites that are off-limits and can only be designated by a bounding box.

Such boundaries are not exclusive to indigenous peoples. In dealing with species at risk, a common occurrence on our fragile planet, there is a need to indicate general but not exact location. This is done to avoid having development proponents remove valued species that might impede development and eco-tourists going to a location and creating more of a problem. At the same time agencies that have access to restricted species at risk data for the purpose of protection of such species and their habitats need to be able to share such restricted data through networked community of practice applications. Similarly, social agencies need to access restricted statistical data to manage and rationalize funded NGO and agency services effectively.

NGOs can be valuable sources of open data as described in Section 2.6. Bringing data from several NGOs together can help in identifying ways to provide better services.

All examples described in this section have geospatial components, that is, they all require maps to gather and display data. Thus open maps should be part of the open data initiative. Google maps or open street maps [6] are not enough in many cases; such maps need to provide multiple layers of data that are not just available with street maps. Far more detail is required as shown by the examples found in [7].

3 Issues Around Open Data

A number of issues arise from consideration of the examples in Section 2. This section categorizes and expands on those issues, which include:

- ensuring that all data that should be open is open
- finding and accessing open data
- providing the right tools to use open data

- keeping open data current
- ensuring privacy of individuals and property
- capturing open data sources
- supporting data redundancy
- sustaining the cost of storage, delivery and maintenance of open data
- governance of restricted data that is open for the purposes of a secure community of practice application

3.1 Ensuring Appropriate Data is Open

Many governments at all levels are committed to open data. How can we ensure that all data generation funded directly or indirectly by the public is also open and accessible? First all government grants and contracts that generate data should be open. Grants can force data to be open by making it a condition of the grant. Similarly contracts for businesses such as consultants should specify that all data generated from the use of government funds or to satisfy government regulation should be open. Data generated by a company with no government funding or mandate would not fall within these categories even if the open data was combined in new ways.

Data to be open must be machine readable and be stored and maintained so it is accessible to all potential users. Storage means two things:

1) Storage on some medium connected to the Internet. This might be supplied directly or indirectly by a federal, provincial/state or municipal government. One such example that is similar is the cloud service called DAIR [8] supplied by CANARIE, Canada's Advanced Research and Innovation Network, primarily supported by the Canadian government.
2) The open dataset should also be registered so that it can be located. Section 3.2 indicates how data might be registered and accessed and what tools might be valuable to manage both the registry and the open data.

Data produced by non-governmental organizations (NGOs) should also be open. Such NGOs should be encouraged to make their data machine readable and accessible as they often produce data that is extremely valuable in a broad societal context. Of course if an NGO receives government funding it should be a condition of the funding unless a significant counter argument can be made.

3.2 Finding and Accessing Data

Although open data can be made accessible; how can it be found? What is needed is a way of locating that data easily and ensuring that it is up to date in a meaningful way. For example, weather data normally must be available in near real-time, whereas soil composition in a given area is fairly constant and may not be updated for years. One could imagine certain data as having a "best-before-date" similar to perishable goods in a retail store.

We propose an open data registry infrastructure that will consist of a collection of of open data registries perhaps that might be organized in a hierarchy by world, country, province/state, municipality and NGOs[3]. A registry will point to open data for its particular part of the hierarchy as well as to its parent, child and sibling registries. Of course since most open data is generated within a specific government department such a structure should work. However, how does one handle data that spans government departments or even jurisdictions such as watersheds that cross state/provincial or even national boundaries. This information could be in the national or world registry. The registry will contain information about the specific open data it references such as metadata, methods of access, provenance, and "best-before-date."

Each open dataset will have to be described by multi-lingual keywords at least based on the United Nations Official Languages.[4] In fact some form of ontology might be made available over time to assist in searching the keywords. Such an ontology might start with a fixed set of words and relations but may be made to learn common words and phrases that are used to describe datasets.

3.2.1 *An Open Data Registry Infrastructure*

An Open Data Registry Infrastructure has two types of users.

1) Data providers: interested parties, such as government agencies, researchers or NGOs make their data available and accessible (by providing information such as metadata, provenance, key words and "best before date" in the open data registry platform).
2) Data users: interested parties such as university researchers, government agencies or commercial enterprises can query the Open Data Registry, find and access the data, and use as required.

An Open Data Registry Infrastructure will include a number of tools including:

1) A register function to allow data providers to supply location of datasets, related metadata, key words, and other provenance data.
2) A maintenance function that will allow the location of datasets, metadata, keywords and provenance to be revised over time. Such a function may be partially automated through tools such as bots.
3) A search function and evolving ontology that allows users to find open data that has been reg-

[3] The registries might also be organized as entity-relations or graphs depending on their use.

[4] Arabic, Chinese, English, French, Russian and Spanish

istered related to specific topics such as; water, watersheds, weather, bio-monitoring, etc.,

4) One or more gateways that support accessing open data using the registry information and composing the open data into useable formats. Standards will be followed where they exist.
5) Tools to support the building of gateways that are likely to be domain dependent.
6) Examples on how to use the Open Data Registry and the open data sets.
7) Sample application interfaces (API) to the registry and accompanying data.
8) User access and controls - a user must sign on to use the registry.
9) Both analytics on use of data sets and user-specific usage, thereby providing knowledge on how the system is used (breadth and depth) and supporting platform management reporting.
10) Users of the Open Data Registry will be able to point the registry at open data that has been derived from previous registry references.
11) The Open Data Registry will also support access to applications that have used the registry successfully.

The GeoConnections Discovery Portal [7] is an initial attempt at providing the functionality of an Open Data Registry, although it primarily relies on a manual approach to finding geo-spatial data. Eventually it is expected that an open data system will evolve in a manner similar to the Internet or Web.

3.3 Using Open Data

With governments at all levels in many countries committed to open data, domain experts will be anxious to explore this data in creative ways to gain insights into areas such as health care, population health, energy, environment and economics to name a few examples. There will be many opportunities to explore the open data and ask what-if questions. Thus data will have to be more accessible [9].

Often the questions will not be well-defined and so requirements will be "made up as we move along." This approach either means that the domain expert will have to be accompanied by a programmer who will react to every whim as the exploratory analysis proceeds or we will require a whole new set of software tools directed toward the domain expert rather than the programmer.

These software tools will be much like the general purpose spreadsheet but be broader in scope including easy-to-use mapping, reports as well as the usual statistical functions and charts. We are sure that the tools will be even broader than we have just described and we have created technologies and a toolkit that takes a step in this direction [10].

Creating these technologies and tools will free up programmers to work on the more difficult problems such as building modeling tools specified by the domain experts and making it easier to access databases that rely on the relational or graph model.

3.4 Sustaining Open Data

Currently there is a flurry to release open data at all levels of government in many different countries and there is talk that open data is the next natural resource due for exploitation. Although this statement has some semblance of truth there are many issues that are not being recognized.

Releasing open data is not a one-time activity; open data must be kept current. Of course currency is dependent on the type of data and its use.

Predicting the weather, which is a current use of open data requires frequent updates and of course large amounts of money for monitoring and capturing the various weather parameters. In contrast characterizing ground cover needs to be done less frequently and only needs to be updated when a catastrophic event such as a severe storm occurs or the land is developed for habitation and "paved over."

Thus, government agencies and other organizations that are providing open data must budget appropriately for open data maintenance. NGOs that are often "poor" may require users of their data to register and pay a fee based on their use of the data. For example, if a company is generating revenue from NGO-supplied open data then a percentage of that revenue should be paid to the NGO. In order to perform this exercise they must be aware of several factors such as:

- What open data is being used? This can be assessed through data analytics available through the registry infrastructure.
- How is the open data being used? Is it being used in a near real-time situation such as weather prediction or for evaluating some aspect of the health system such as the efficacy of the delivery of health care?
- Is the open data changing?

By knowing what open data is being used, how it is being used and whether the data is changing, the focus can be on keeping that data as current as necessary. There may be times when open data needs to be updated because of a new use. In this case there must be a protocol in place that supports requests and is responsive.

Governments are devoting significant resources to publishing open data and exclaiming over its potential value [2]. The Canadian federal government alone has announced the release of at least 200,000 data sets and is heavily promoting open data and transparency at all levels of government [3]. The members of the G8 have also adopted an open data policy [11].

The flurry of activity has certainly generated a large amount of interest in the social and economic value of open data and its potential in making government more transparent. How will the government sustain the level of activity needed to make data open, available and current? What mechanisms can be put in place? A two-part scheme comes to mind.

1) The use of open data can be monitored as to use; open data not being actively used can be taken out of circulation or at least not kept current.
2) Open data can be requested through a "freedom-of-information" request with appropriate mechanisms in place to ensure quick response. In this way, data not being kept current can be revived and data corresponding to new uses can be made live.

3.5 Privacy of Individuals and Property

Releasing open data can have serious effects for society. Privacy of individuals can be affected where identity can be inferred from data. A classic example of inferring identity based on key words released by an Internet service supplier can be found in [12].

There are also issues around the protection of societies or natural phenomena. For example, disclosing information about areas important to the First Nations as mentioned in Section 2.5 or about natural species at risk may attract the curious and be detrimental to areas that are already fragile.

Such an approach can be particularly difficult when the information is displayed on a map. Possible methods of making this information less "public" are:

1) creating a buffer zone around the information so that its location is not too specific.
2) indicating that there is something of importance within a buffer zone but not indicating the exact contents.

If the scientific or similar community needs access to more accurate information, then that can be provided based on suitable guarantees.

3.6 Data Sources

Releasing open data raises many questions about the data sources. The next few paragraphs explore some of these issues.

Currently the UK government is planning on using a five-point scale originally conceived by Tim Berners-Lee [2] and aiming for level 3, which is described as making the data available in an open non-proprietary format such as those that use comma or character-separated variables (CSV) or XML.

Normally CSV implies a flat file. Will the XML version be used to describe flat files or will the structure of a relational, object-oriented or graph databases also be encoded in the XML description? If the latter is the case what metadata will be provided? It seems that relationships among data objects are often as important as the objects themselves.

How granular and raw should open data be? For example, min and max temperature data should be published at least daily, although if one is looking for variations over the day more frequent values might be necessary. Data might not be published in a completely raw format as corrections might have to be applied owing to failure of specific sensors.

A second example illustrates the complexity of deciding how granular data should be. A country is collecting information about foreigners entering during the year. How do they publish the data, by country of origin and how frequently, by day, month or year. It may be noticed that people from a certain country or area of the world may be entering during a certain period. Thus, based on the analysis of this data, a tourism bureau may target this area for advertising just before these periods. Obviously the granularity of the data depends very much on the ingenuity of the application developers in using the data. Since this data must be collected on a daily basis and must identify the origin of the traveler, it is certainly possible to provide data in any of the forms just described.

Certain data such as weather data is collected and distributed in almost real-time to support weather forecasting. However, there is other data that is collected in near real-time that is extremely valuable. In Section 2.1 we refer to watershed modeling. If this model could be operated to accept data from sensors measuring rainfall or snow-melt then it would be possible to predict and respond in almost real-time to the potential impacts of major weather events that can cause flooding and serious infrastructure damage and possible loss of life. Such data is often collected by local authorities and not by national governments. Even if national governments do collect such data, by the time the data reaches them, the crisis has passed and the damage is done.

Crowd-sourced data can be as valuable as that delivered by government. The Invasive Tracking System described in Section 2.3 is one example of the general public using data and digital photographs to identify and report on invasive species. In fact this is the only way to get such information as most levels of government are not able to afford the manpower for such a task. The question then becomes one of how to incorporate crowd-sourced data into the open data strategy and how to register availability of such data.

It is often stated that if data is not available or not available in the desired form, then the required data can be acquired through a freedom-of-information request [2]. How practical is such a request? Will the data be released in a timely fashion? How will such requests be prioritized? What kind of tools will be needed in order to satisfy such requests? There are many such questions that need to be addressed.

3.7 Data Redundancy

In several of the examples in Section 2 there is a need to copy open data for performance reasons. Such a situation is particularly important when open data is being used in modeling. One does not want to unpack data for each run of a modeling program as such a function is just too labor- and time-intensive.

One must devise tools that allow downloading, unpacking and storage of data and that can be quickly configured for new situations. In addition, there must be mechanisms in place that notify the data user about updates to the data so that they can be downloaded when appropriate.

3.8 Sustaining the Cost of Storage, Delivery and Maintenance

Federal and provincial/state governments normally have the capacity, longevity and resources to act as repositories of data important to the public interest. However, valuable open data also resides with smaller governments such as municipalities, NGOs, and universities and colleges. For example, municipal governments, NGOs and universities produce valuable land use data, social data and scientific data respectively. To be even more specific, social agencies in large cities overlap in both function and territory served. Knowledge of that information in the form of open data could help to rationalize services and possibly result in significant cost savings.

Although these organizations are important producers of valuable open data they often do not have the capacity to maintain an open data repository and frequently seek public funding to support their open data mandate. Of course public funding either from the public, business or government does not always correspond to the needs of the organizations.

Such considerations about open data raise a number of questions that need to be answered including:

- How is the data gathered and who pays for gathering it?
- How is this data stored, delivered and who pays the costs of these services?
- Who shoulders the cost of maintaining the data once it is collected the first time?

Such questions raise the concept of examining a multi-sectoral approach that would satisfy the requirements of all participants. However, upper level governments would have a special fiduciary responsibility to operate and maintain the system for open data. Storage, maintenance and delivery of open data would be managed by a group representing all sectors and would determine all the participants and the value of the open data they would add. Such valuation could occur through the use of analytics of data use over time.

4 Creating Value from Open Data

The literature and pronouncements about open data besides supporting government transparency indicate significant economic potential in two areas. The first is improved public service at reduced cost and the second involves accessing and combining data in new ways to create value for both society and business.

There are certainly strong indications that economic benefit might accrue to businesses working with open data based on the experience of the open software movement. Here companies such as Red Hat, Ubuntu and IBM have profited incredibly from the managed distribution of the Linux open source operating system. Other companies are attempting similar things with products such as the database MySQL.

At least two prominent working papers [2], [1] present examples of how the benefits of open data can save public money directly or indirectly. They also cite examples of businesses and services that have been created [2], [1].

In [1] they cite California and Texas as saving millions of dollars through a transparency web site, while Canada is using similar approaches for fraud detection. This same document shows how access to open health data can increase competition among hospitals to provide better services through reduction in MRSA infection rates.

The presentation in [2] outlines businesses and services that have evolved from use of open data. They range from information about disruptions on the London Underground to a Contracts Finder similar to the MERX [13] system in Canada which has been operational for over 10 years. The difference between MERX and the Contracts Finder is that the Finder will have an open interface thus allowing the potential development of tools to support new services to access the contracts database.

Although we are seeing opportunities for the use of open data, the authors feel we are only seeing the tip of the iceberg. How can we start to unlock the real potential of open data?

Some propose a hackathon approach [14] in which computer programmers and others involved in software development collaborate intensively on software projects. Is such an approach really adequate when working with open data? Rarely do software developers have the domain experience or depth of knowledge to identify the type of problems described in Section 2. When a new "app" is produced it is often based on the limited domain experience of the software developers rather than based on needs that have been identified by domain experts.

In fact this mirrors one of the biggest problems that faces software development to this day. We have highly skilled individuals who can create powerful software using a variety of tools, but they often do not understand the intricacies of the domain they are modeling. This has led to more and more effort being put into requirements engineering and to creation of tools and processes such as agile software development [15] and software transparency [16] that support close interaction between the craftsman (programmer) and the domain expert.

We propose a different model, one that has worked in many of the applications developed in Section 2. Rather than just forming teams of software developers we need to form a mentoring network that combines domain experts and software developers into a coherent team [10]. Domain experts could post problems on a bidding site much like the one described for contracts earlier in this Section or work through a group of colleagues to put these problems in front of the software development group. It must be clear that problems are not to be posed and then the experts walk away. Rather they must mentor and guide the software development team as they progress toward a solution. A similar approach seems to have been taken by the Open Development Technology Alliance [17].

5 Related Work

Open data as described in Section 1 like big data has become an important new direction in information technology. Governments are releasing data so as to be more transparent to their constituents, namely the general public and business [1], [7], [2], [3], [11]. Not only are they claiming to be more transparent by releasing data but there are also statements about the value of the data both in saving money delivering government services and in new businesses delivering economic and social value.

Most organizations outside government that are focusing on open data [1], [18], [19] with one exception [17] seem to focus on various types of education. The education seems to be toward "selling" the value of open data and showing governments at all levels how to construct an open data policy. The exception appears to be the Open Development Technology Alliance [17] and our own research team that is working toward an agenda that is at least demonstrating through multiple practical applications how open data might be used. Except for a recent conference [20] there appears little attempt to address all the research issues around open data as suggested in this paper.

6 Conclusions

The open data sector appears to be driven by the need of government to be more transparent to its constituents. Open data is likely to create interesting opportunities that could deliver not only transparent government service but benefits such as improved government service and societal and economic opportunities.

However, dealing with open data is far more complex than appears on the surface. Data will need to be combined from all levels of government and with data from NGOs, businesses and other sources such as university research labs. Maps will also be important as most open data has a geo-spatial component. There are many issues relating to data sustainability/maintainability and privacy that are very important.

This paper is an attempt to outline many of the issues associated with open data and to suggest the beginnings of an open data research agenda.

Acknowledgments

The authors thank Bob McFettridge, Executive Secretary of the Federal Biodiversity Information Partnership, Doug Mulholland, a senior member of UWCSG, Geoff Bellew of Communitech and Carlos Lucena of the Pontifical Catholic University of Rio de Janeiro (PUC-Rio) for their insightful comments and suggestions for content and additions to this paper.

References

[1] "Open Data Institute Business Plan," http://theodi.org/about-us, 2012.

[2] "Open Data White Paper Unleashing the Potential," https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/78946/CM8353_acc.pdf, 2012.

[3] "data.gc.ca," http://data.gc.ca/, 2013.

[4] S. Lawrence and P. Macklem, "From Consultation to Reconciliation: Aboriginal Rights and the Crowns Duty to Consult," *Canadian Bar Review*, vol. 79, pp. 252–279, 2000.

[5] M. Asch and P. Macklem, "Indigenous Rights and Canadian Sovereignty: An Essay on R. v. Sparrow," *Alberta Law Review*, vol. 29, no. 2, pp. 498–517, 1991.

[6] "Open Street Maps," http://www.openstreetmap.org/#map =5/51.500/-0.100.

[7] "GeoConnections Discovery Portal," http://geodiscover.cgdi.ca/.

[8] "CANARIE - Digital Accelerator for Innovation and Research," http://canarie.ca/en/dair-program/about.

[9] A. Koller, "Opening Open Data," Opening Open Data http://www.w3.org/2013/04/odw/papers, 2013.

[10] D. Cowan, P. Alencar, F. McGarry, and C. Lucena, "A Web-based Framework for Collaborative Innovation," David R. Cheriton School of Computer Science, University of Waterloo, Tech. Rep. CS-2012-02, 2012.

[11] "G8 Open Data Charter," https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207772/Open_Data_Charter.pdf, 2013.

[12] N. Carr, *The Big Switch Rewiring the World from Edison to Google*. W. W. Norton and Company, 2013.

[13] "MERX E-Tendering Suite," http://marketing.merx.com/Resources/MERX_Private_Tenders.pdf.

[14] "Hackathon," http://en.wikipedia.org/wiki/Hackathon.

[15] "Agile Software Development," http://en.wikipedia.org/wiki/Agile_software_development.

[16] J. C. S. do Prado Leite and C. Cappelli, "Software Transparency," *Business & Information Systems Engineering*, no. 3, pp. 127–139, 2010.

[17] "Open Development Technology Alliance," http://odta.net/2976/home.

[18] "Canadian Open Data Institute," http://opendatainstitute.ca/, 2013.

[19] "US Open Data Institute," http://flowingdata.com/2013/10/29/u-s-open-data-institute/, 2013.

[20] "Open Data on the Web," http://www.w3.org/2013/04/odw/, 2013.