White Paper:
A Data Commons in the Exchange Space

Greg Kidd,
Founder and CEO
Karen Gifford, President

September 16, 2010

A Data Commons in the Exchange Space

The term "Data Commons" refers to a cloud of data available for syndication. In this paper, we discuss some of the issues surrounding the establishment of a Data Commons for data related specifically to the exchange of goods, services and information—including commerce-related data.

Where there are many content sources and many content users, as in the exchange space, a clearing system that will normalize and index all content will help consumers and data users more efficiently than the discrete, bilateral transfers of exchange-related data between two parties that exist today. We envision the Data Commons as a collaborative effort between a number of for-profit and non-profit actors.

I. Why A Data Commons?

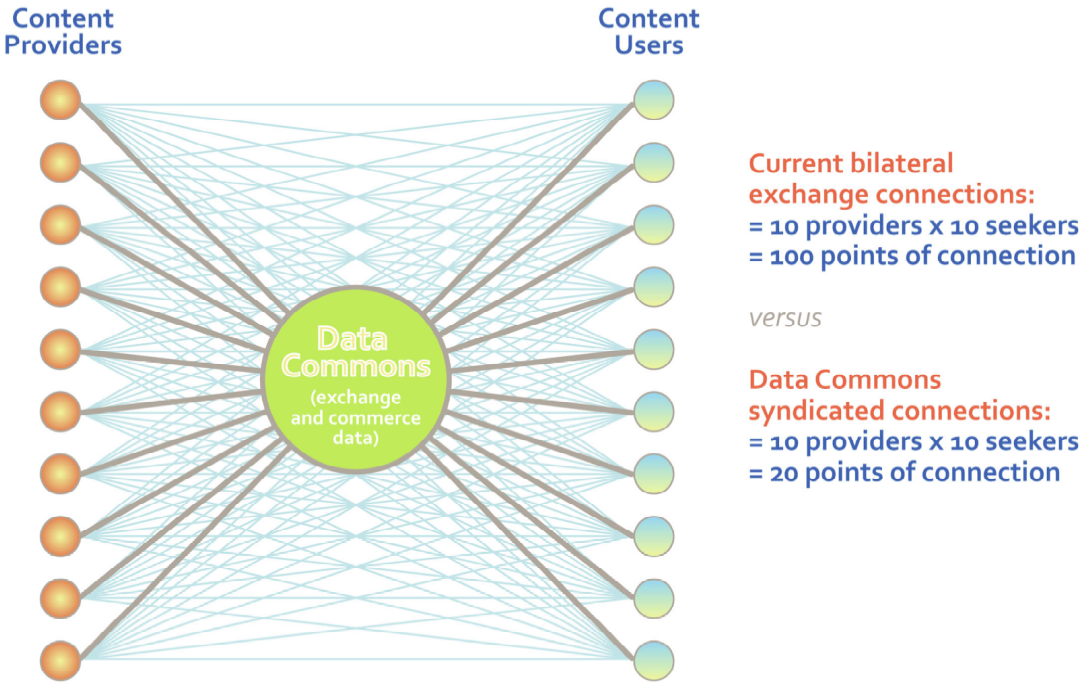
Potential Efficiencies of a Data Commons

To envision the potential value of a Data Commons, the payment system operating in the United States provides a useful analogy. In theory, banks could clear checks and electronic payments between one another. Because there are 10,000+ banks in the U.S., direct bilateral transfers could result in 10,000 x 10,000 (100 million) points of connection to achieve settlement of all obligations. This unwieldy situation is precisely what exists with respect to exchange-related data on the Internet. Most of what we call ***postings***—transaction-specific information about individual items offered for exchange, such as a watch for sale or a job offering—are found only by a search on the individual websites where they live. To do a comprehensive search for a particular item one might wish to buy one must visit numerous separate sites.

The United States banking industry has evolved a much more efficient system. Instead of requiring each bank to contact directly all other banks with which it has settlement transactions, the Federal Reserve operates a clearing and settlement system, taking inbound bulk payments and making outbound bulk payments from/to each bank. This arrangement results in 10,000 x 2 (20,000) points of connection—a 99.98% efficiency improvement over a bilateral settlement system. One can argue whether such a clearing and settlement function should be in the private or public domain, yet there is little argument over the benefits of efficiency.

The illustration below depicts an example the efficiencies that could be achieved if a Data Commons were in place to act as a central clearing system for postings in the exchange space on the Internet, si milar to the role played by the Federal Reserve in the banking industry.[1] Here, the same ten content users access data from ten content providers through many fewer points of connection:

[1] For the Data Commons, clearing always refers to the most efficient sourcing and delivery of posted data between seekers and providers. Settlement refers to situation where an economic rent obligation also travels with a particular piece of posted data. In such cases, seekers and providers of posted data encumbered with settlement obligations can either choose to conduct settlement bilaterally (even if the data flowed through the Commons) or may opt to utilize settlement options directly through the Data Commons when such financial infrastructure exists.



Potential for Increased Value Creation

A Data Commons offers the possibility of greatly expanding the value creation that takes place on the Internet. A large portion of that value involves the ability to "discover" content—either by traditional searches (Google or Bing being the leaders) or, increasingly, by referral (Twitter & Facebook being the emergent leaders).

In the commerce space, a robust market in advertising and affiliate sales that rewards leads (both in the form of click-throughs and actual sales) has led to a monetization model that allows one set of websites to be rewarded for bringing qualified leads to another set of sites that might benefit from the increased traffic. The evolution of this relationship has led to increased innovation and experimentation between sites that are good at gathering content and those that are good at using it (and everything in-between).[2]

Currently, sites that seek to identify and aggregate exchange-related postings—those likely to lead to click-through and sales payment rewards—must approach (on technical, legal and business terms) a large number of separate sites. Just as for U.S. banks, a centralized clearing system, or Data Commons, would greatly reduce time and operating costs, opening the arena to both small and large operators. It is worth noting that in so doing, the Data Commons would play a democratizing role in the exchange space.

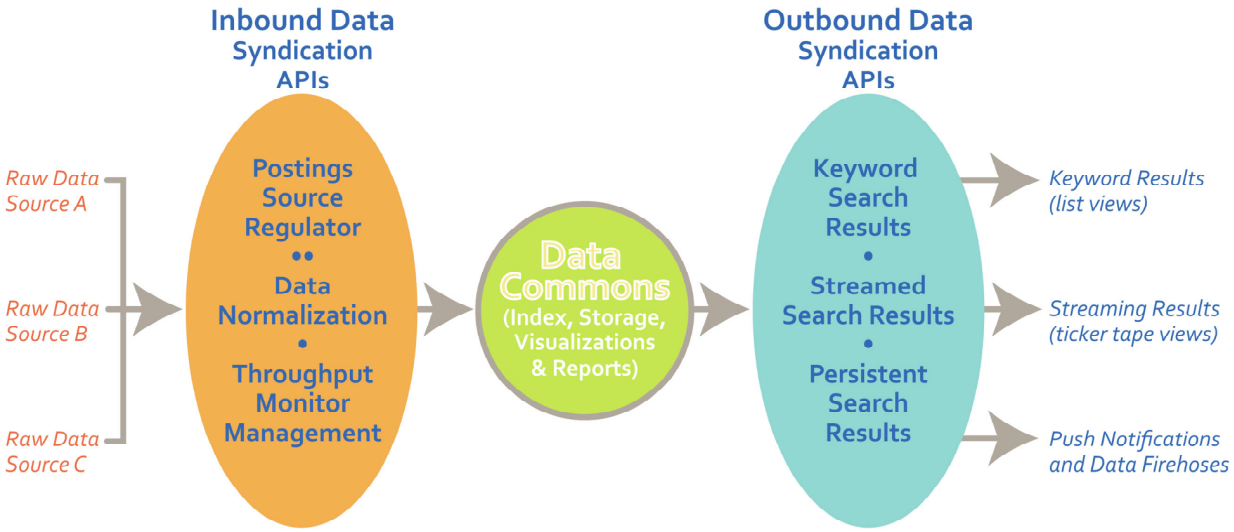
[2] Sites hosting content sometimes pay for such referrals and affiliate sales. Sites wishing to benefit from such payments seek to gather the underlying data that will generate the highest probability of click-throughs or sales with the least friction. Most of Google's profits derive from optimizing click-throughs on its own and third party sites; and Amazon's profits too are likewise substantially generated by garnering (and paying for) affiliate sales from third parties who may have users but little or no sellable content of their own. Some other websites diligently attempt to keep all their activity and lead generation in-site (i.e. Craigslist) by refusing to pay for referrals or share sales proceeds. However, it is worth noting that almost all commerce-based sites, with the exception of private exchanges, post the facts about their offerings squarely in the public view.

In addition, by substantially increasing the efficiency of content discovery, a Data Commons would greatly expand the type and amount of data accessible to businesses and non-profits interested in making use of it. More efficient clearing of information between seekers and providers of data in the commercial world would be expected to lower the cost of establishing affiliate sales and other direct marketing programs. Outside the commercial realm, increased efficiencies and lower costs could facilitate exchange-related projects that simply were not feasible in the past.

II. What Might a Data Commons Look Like?

In this section we discuss what a Data Commons might look like as a practical matter. We envision a Data Commons as much more than a simple repository. A Data Commons that includes such functions as normalization of raw would facilitate meaningful search across many hundreds data of sources at once. Because the Commons would hold such data, it could also support any number of applications for outbound products such as highly customized data streams. The illustration below shows a possible model for a Data Commons:

Illustration 2
Data Commons Model



In this example, the Data Commons would effectively be bounded on the incoming and outgoing sides by application programming interfaces (APIs). The APIs would define interactions between entities outside the Commons and the Commons itself. We anticipate that the following APIs would facilitate the efficient management of raw data collected by the Commons:

- **Posting API** – to regulate quality of incoming raw data postings, e.g., confirming data sources and keeping redundant postings from entry to the Commons;

- **Normalization API --** to set standards for incoming data normalization[3], including geo-codes, text categorization and hashtags, annotations[4] (such as images, etc.), and to add meta-data (such as source descriptions, date-time stamp received, etc.) associated with the postings; and

- **Monitoring API** – to permit those managing the Commons to review information on throughput into the Commons and data queue behaviors that take place en route into the Commons.

These inbound APIs would support achievement of a principal objective of the Data Commons: to index, store and report on real time updates of similar types of data originated from many diverse sources. This repository of normalized data would be useful to a wide variety of users in the commercial, academic and non-profit spheres, and would afford a substantial improvement in the efficiency with which exchange-related data can be accessed.

The following APIs would provide data in outbound visualization formats, useful for a wide number of potential users:

- **Search API** -- allows users to perform free form search on keywords or structured search on indexed terms; delivers traditional search results in a list view.

- **Stream API** – allows users to view a constant stream of individual postings in real time at a specified rate for any criteria that can be specified in the Search API. The data stream may be a sample of matching postings if the number of match results exceeds the user's specified stream rate, or the stream may contain a mix of real time and historical postings if the number of real time match results are less than the specified stream rate. The Stream API may also be implemented as a widget that can be built into Web and mobile applications where a streaming view of data is both useful and eye-catching.

- **Notification API** – allows users to conduct persistent searches, whereby new match results are pushed to a website or mobile device as they occur. Notifications may be highly selective if based on keyword matches of infrequently used terms, or they may be voluminous "fire hoses" of data if they are based on broad categories or geographies of criteria.

[3] In order to make data comparable across all sources, the Commons employs its own independent categorizing logic and mapping rules to and associate data with a particular metropolitan area where relevant. All inbound postings are mapped based on translation tables and forward and reverse geo-coding mechanisms promulgated by the Commons for use by any submitter.

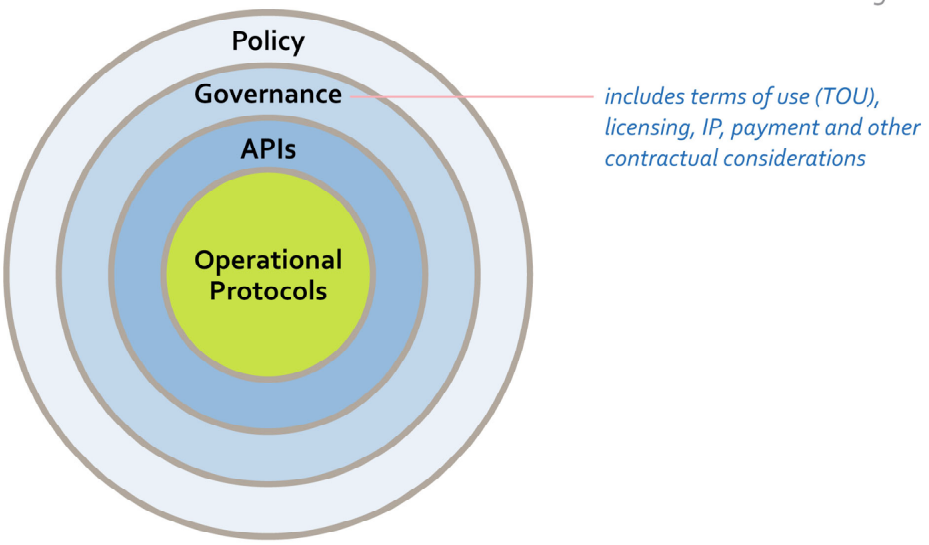
[4] Structured annotations are typically organized as key value pairs such as "brand=Ford" or "model=Thunderbird" or "weight=10lb" that are contextually relevant to some categories of postings

Governance and Operational Elements of a Data Commons

In order to function effectively, a Data Commons for the exchange space on the Internet will require a number of governance and operational elements. In this section we outline some ways in which these organizational elements of the Commons may work together.

Certain entities may best be suited to supply or oversee particular organizational operations. For example, a neutral entity could serve an important role in governance to bring interested parties together, promoting dialogue and articulating common standards that govern the Commons, with an eye for "the common good." Private parties may drive the shape of certain organizational elements. We do not go into any great detail here about how such roles would be defined, but rather hope that these notes may prove the basis for useful discussion.

Illustration 3
Data Commons Organization



Governance – We anticipate that much of the work to establish and maintain the Data Commons can be accomplished without elaborate or formal governance structures. As we point out above, one of the primary benefits of the Data Commons will be the significant efficiency improvements available to those seeking access to mass exchange-related data. In our view, any governance structure for the Commons must be consistent with its technical efficiencies, in that Governance should not seek to impose unnecessary friction on users who want to share and access Commons data.

We believe that governance structures necessary to manage the Commons can be embodied in a relatively small number of standard contractual provisions, most likely set forth in the Commons' terms of use (TOU). We discuss some of the issues these standard contractual provisions will need to address in the following section of this paper.

APIs – As noted in some detail above, we envision that a series of APIs will set the boundaries of the Data Commons, and describe in technical terms the form and means by which data may travel into and out of the Commons.

Operational Protocols – Those managing the Data Commons will need to develop and maintain a number of operational protocols to support the smooth throughput of data into and out of the Commons. These protocols should include quality assurance activities and other activities related to generating and managing meta-data (e.g., data source information versus information about the data itself).[5]

In addition, a successful Data Commons must "operationalize" requirements for complying with its legal and contractual frameworks. (We discuss in more detail below issues that may call for specialized contractual provisions.) For example, those managing the Commons will undoubtedly need to develop protocols to prevent disclosure of certain types of sensitive data, and tracking and other activities to support correct payments associated with other specific types of data. One specific example of a legal obligation that must be operationalized is the Digital Millennium Copyright Act (DCMA) takedown responsibilities that will arise for the Data Commons when data is mistakenly submitted or misclassified in terms of copyright.

It is worth noting that while intellectual property and confidentiality concerns place a burden of responsibility on the Data Commons to limit some syndication requests, the inverse is also true: in dealing with publicly available, factual data, the Data Commons has a responsibility to assure that such data is free and fairly accessible to all, on an equal basis (i.e. without discriminatory permission-based interference by any third party).

Considerations for the Governance Framework: Data-Specific Issues

As we note above, developing a limited number of standardized, data-specific contracts (or licenses, in the case of data with associated intellectual property rights) is a key element in the establishment of a Data Commons. These data-specific contracts are of particular importance because they will permit the smooth function and operations of a Data Commons from a governance and legal standpoint—an aspect of the Commons that is at least as important as the technical elements that allow easy access to the Commons' data.

To minimize transaction costs associated with Data Commons operations, the contractual framework supporting the Commons must define the obligations of the various parties in a clear manner, with a minimum number of specialized provisions. At the same time, this contractual framework should adequately address legitimate, data-specific issues so that data providers and users do not feel the need to create side agreements that could impinge on the Commons' effectiveness.

Those developing the contractual framework must bear in mind that every specialized contractual provision will result in the development of corresponding provisions in Commons' APIs and operational protocols. Identifying those issues that legitimately merit specialized contractual terms at the outset can help avoid difficulties in the future. This is so even if the

[5] The Data Commons itself creates no original data, but the Data Commons may create much derivative data through efforts to normalize and annotate data across sources and to enrich the data with meta-data.

Commons does not initially support all types of exchange-related data. The Commons can better take into account progressive expansions of its universe of data if the accommodations needed to support new types of data are considered in advance.

We identify four broad categories of data that have implications for the contractual framework necessary to support the commons:

- Copyrighted and patented works that carry strong *intellectual property rights* and responsibilities;

- Confidential and otherwise *private data* where either the identity of the sender needs to be kept anonymous and/or the identity and certain other details are only authorized for delivery to particular users of such content; and

- Facts, and opinions about facts, *publicly available* and in the *public domain* that are subject to broad protections under the First Amendment right to free speech (the free exchange of factual information in the public domain may also be subject other legal protection, including that of antitrust law); and

- Data associated with a right to payment of one kind or another, which we refer to here as *economic rights.* A significant component of this data consists of postings from commercial websites that have established advertising or affiliate sale programs. Commercial websites may also wish to establish standard terms for third parties to use their data as part of Commons programs, and could devise ways to make use of the Data Commons to distribute their terms of use (TOU) and attendant data tracking.

The matrices below provide one view of how these data categories interrelate for purposes of developing appropriate contractual provisions.

Permission and Payment - The first matrix depicts the interrelationship of intellectual property (IP) rights and economic rights (ER) associated with data moving through the Commons:

	IP Rights	Economic Rights
Facts	No permission required	Payment at discretion of source, independent of permission to use
Creative Works	Permission at discretion of source	Payment at discretion of source, tied to economic rights

As the matrix makes clear, it is necessary to distinguish *permission* to use data that may have IP rights associated with it from an ER *right to payment* that might be associated with that permission, or which might arise when such data is used as part of a direct advertising program. Developing standard contractual provisions that govern permissions and payments will greatly enhance efficiencies associated with use of the Data Commons. It is worth noting here that purely factual data that is publicly available may move freely into and out of the Commons, without the need for specialized contractual provisions.

Privacy and Confidentiality —This matrix shows the interrelationship between various privacy and confidentiality rights that could be associated with data moving through the Commons. Here we envision situations in which information coming into the Commons may be partially or entirely unavailable to the public view.

	Public Content	Private Content
Signed	All users can see content and identity of contributor	Specified parties can see content and identity of contributor
Semi-anonymous	All users can see content; Commons (or intermediary) cloaks identity of contributor	Specified parties can see content: Commons (or intermediary) cloaks identity of contributor
Anonymous	Anyone can see content; identity of contributor unknown	Specified parties can see content: identity of contributor unknown

We envision that such situations could arise in a variety of contexts, a few examples of which we list below:

- The data is subject to claims of privacy, e.g., personally identifiable medical information (such data may also be subject to various legal and regulatory regimes);
- A buyer or seller chooses to remain anonymous in a commercial transaction;
- A whistleblower seeks to anonymously report behavior that may violate the law; and/or
- The victim of a crime seeks emergency response.

As can be seen from these scenarios, the need for privacy or confidentiality may arise for a variety of reasons, which will call for a number of different contractual provisions and related operational support.

III. Conclusion

A Data Commons would greatly increase efficiency of access to exchange-related data. We believe that these efficiencies, and the attendant decrease in costs and democratization of access, well justify the work involved in establishing such a Commons.

The ideas outlined above about how a Data Commons might be structured represent our initial thoughts on what will undoubtedly be a complex and multi-stage project. We hope that we have given readers food for thought, and welcome a dialogue about whether and how to proceed.