A beginner's guide to
OPEN DATA

Open Data.

Relatively unexplored and full of possibilities, a vast collection of data exists online for anyone to access. The value that open data provides goes beyond community participation and improved products. Having data readily available regarding taxes, housing prices, crime, navigation, energy efficiency, and more, opens up a whole new level of global transparency and accountability.

"Open data and content can be freely used, modified, and shared by anyone for any purpose"

- opendefinition.org

This eBook will take you through the ins and outs of open data. You will see how to avoid common pitfalls and instead effectively share your information with the world.

Who Produces
Open Data?

All levels of Government

Governments are currently the main providers of open data and have been the forefront of the open data movement. The reason? Firstly, in a drive to increase transparency and accountability. Secondly, they have realized the benefits citizens get from innovative applications and services that have been built by third-parties on top of the data.

Almost all western governments are now mandated (e.g. INSPIRE) to provide open data to their citizens, and some emerging economies (e.g. Kenya and Moldova) are going down the same path. Data is provided at all levels of government in Europe and the USA (e.g. city, state and federal). In emerging economies open data can help combat corruption and poverty by promoting transparency and empowering citizens.

As we move towards smart cities and gain the ability to measure and control almost any physical object, data volumes are going to explode. This means that the importance of keeping a low cost of entry for developers is critical.

Flagship implementations are data.gov and data.gov.uk.

Non-Governmental Organizations

NGOs have always paid attention to the democratization of data because they are often working for the provision of better public services or planning development projects and open data can directly help them achieve their goals. They are now also starting to produce open data themselves.

In 2009, the Hewlett Foundation and the Gates Foundation began sharing where they spent their aid money.

Today, over 341 organizations have published data about their operations and spending via the International Aid Transparency Initiative (IATI)

Relief efforts can also benefit significantly from open data. Crowd sourcing and geomapping can ensure that different organizations are not spending time and money acquiring the same data and that the development and aid money goes where it's needed most.

Open Data present in Nepal

Within 48 hours after the devastating Nepal earthquake, 4,000 mappers had mapped out 13,199 miles of road and 110,681 buildings. These maps allowed aid groups to make rescue plans and find the safest and fastest routes to access people.

Open Knowledge, Open Aid and Open Governance … transform development and hold greater hope for the problems [in developing countries]."

- Sanjay Pradhan
VP World Bank Institute

TED

VIEW THE TED TALK

Academic Institutions

Academic institutions have recently started opening up their data. The University of Oxford and the University of Southampton both provide information on courses, food services, events and news. But the real power of open data in Academia comes from making *research* data open.

Fuelled by the success of sharing data on Alzheimer's, there is a movement beginning for more transparent research practices. As there are a lot of complexities around sharing research data, progress is slow. Despite this, more and more research funding agencies and academic publishers are supporting and even mandating data sharing.

Openly accessible research data is also at the heart of the EU's €80 billion Horizon 2020 program, which gives further hope of dramatic progress over the next 5 years.

Open Data Makes History

One of the greatest successes of open data was The Human Genome Project. In this feat, the human genomic sequence information was made public. As a result, a genome can now be mapped in a few hours, costing less than $1000.

Private Companies

Even though governments are the main providers of open data, there is a significant proportion of corporations who are starting to realize that producing open datasets can improve their bottom line. Benjamin Herzberg of the World Bank Institute calls this new frontier The Open Private Sector.

"**For many companies, openness of data can translate into more efficient internal governance frameworks, enhanced feedback from workers and employees, improved traceability of supply chains, accountability to end consumers, and better service and product delivery."**

- Excerpt from: The Next Frontier for Open Data: An Open Private Sector

VIEW THE FULL ARTICLE

Why does it matter?

The debate over whether or not this data should be made available to the public is still ongoing. Traditionally, many governments have viewed the sale of data as a revenue stream. Other governments view the infrastructure costs of freely opening data as prohibitive. Lastly, the raw data may not be in a form that is easily shared.

As data lovers, we at Safe Software believe that data should never be locked inside applications or formats. Data should be free to use whenever, wherever, and however it's needed.

By opening data, power can be handed from the government to citizens. These citizens then have the option to examine the data and answer questions they may have. Researchers and journalists who want to gather and analyze the data will be able to tell stronger stories, and developers can use the data to build applications.

Access to open data means innovation.

FREE YOUR DATA

Be Responsible for Quality

ENSURE YOUR DATA'S QUALITY BEFORE IT'S MADE PUBLIC.

Making decisions based on bad data could have extreme ramifications. Before sharing a dataset to the public all aspects should be checked for completeness, correctness, consistency, and compliance.

Good quality data means checking that it fulfills requirements, then repairing it where it doesn't pass.

CSV Trouble?
Use
CSVLint.io

Our [data quality checklist](#) can help you validate and repair your geospatial data. Validation steps will vary to some extent depending on the type of data (2D, GIS, raster, etc.), but this list will provide a good guideline. You can use manual verification and out-of-the-box tools to verify your data, or you can use FME to [detect and repair problems automatically](#).

Ultimate
Geospatial Data Validation
CHECKLIST

safe.com/qa

Check the schema
☑ Feature type (i.e. layer, level, table ...) names
☑ Attribute names and types
☑ Coordinate system
☑ Allowed geometries

Check the data values
☑ Correct data type for the field
☑ Within the valid range/domain
☑ Check for duplicates (e.g. of a unique key)
☑ Are null values allowed? If so, are the null types consistent (NaN, infinity, empty strings, etc.)?

Validate the geometry
☑ Self-intersections
☑ Degenerate or corrupt geometries
☑ Null geometries
☑ Vertices with missing normals
☑ Texture coordinates, for geometry with a texture
☑ Invalid solid boundaries
☑ Invalid solid voids
☑ Non-planar surfaces
☑ Duplicate consecutive points, in 2D or 3D

Compliance to standards
☑ OGC
☑ INSPIRE
☑ Other international standards or trade standards
☑ Your company's standards

Format-specific QA/QC
☑ CAD data: ensure the robust extraction of layers, geometry, text, line types, blocks, extended entity data
☑ XML / JSON: validate the syntax or schema
☑ Tabular data: ensure values pass logical tests; check integration with spatial details
☑ Databases: check the data and geometry before attempting to load it into a central repository
☑ Point clouds: check for correct components and values

Workflow-based validation
☑ Detect differences in an updated version of the same data
☑ Validate submitted data and immediately give feedback to stop bad data from being processed
☑ Other requirements for your workflow

Repairing and reporting bad data
☑ Map the schema to fit the destination data model
☑ Geometry manipulation
☑ Enforce compliance with your standards
☑ Flag the bad data and return it for human analysis
☑ Measure and describe the quality in a standardized way
☑ Send the report

FME

Formats

If data is published online in an unknown format, will it make an impact?

No.

It's not open data if it can't be easily used by the public.

Data should be both machine readable and human readable. That is why we have one machine-readable format and one human-focused format in both tabular and spatial. For example, XML and GeoJSON are easier for machines to read.

To be sure your data meets its full potential, be sure to share it in these required formats...

Suggested formats:

Tabular	Spatial
CSV JSON	Shapefile GeoJSON

Supplementary formats:

Tabular	Spatial
XML Excel	KML Mapinfo TAB File Geodatabase AutoCAD DXF/DWG

Format Fails

PDFs MAKE FOR POOR OPEN DATA AND GOVERNMENTS SHOULDN'T HAVE TO USE TUMBLR

In a time of crisis, data can be a critical factor to ensure society's safety. In 2011, when Hurricane Irene ripped through New York, the storm's path was closely monitored and mapped. The safest evacuation routes were determined by this mapping and shared with the public.

The problem was, these evacuation routes were displayed on a complicated PDF Map. Besides being overloaded with information and thus complicated to read, the PDF could not be updated easily.

To top off this underwhelming PDF, the city hosted it on Tumblr when their own infrastructure crumbled.
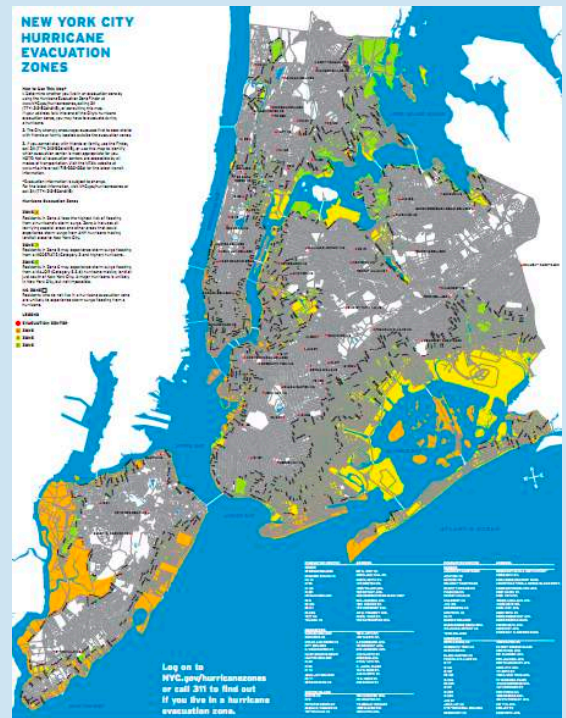
Format Victory

OPEN DATA FORMATS SHOULD ALLOW FOR INNOVATION.

Collision data recorded by the NYPD was originally released as a PDF due to legislation enacted by the council.

The NYPD's apparent motivation behind using a PDF was to limit the public's interaction with the data and reduce the ability to manipulate it.

Frustrated with the inflexibility of the data, an enterprising citizen created the NYPD Crash Data Band-Aid, a tool that scrapes the PDF and turns it into a geocoded CSV. He wanted to be able to use and analyze the data to improve public safety.

The point was made. Since then, the NYPD has moved all of their data to the NYC Open Data Portal. Data can now be downloaded (via Socrata) in machine, and human-readable formats of choice. Innovation was the result and various apps were built on top of the new data at the BIGAPPS NYC 2014 competition.

Process

WHAT TO THINK ABOUT BEFORE SHARING DATA.

Projection Support

For spatial data, users should be able to choose their projection. Local (e.g. State Plane or British National Grid) and global projections should be provided. Recommended Global projections:

- WGS 84 Lat/Lng (EPSG: 4326)
- Spherical Mercator (EPSG: 3857)

Data Should be Updated...

In an eye-opening article done by thomaslevine.com , it can be seen that almost all open datasets are static and not regularly updated.

Here are a few steps you can take to ensure your data is up to date.

Make it an API

Provide your data as a published feed (e.g. RSS) or API rather than downloadable data. That way people can consume the endpoint rather than the actual data. Plus, if you make a change, it will automatically be reflected in their app.

Connect to the Master Database

Rather than duplicating your master data store in your open data platform, connect your open data platform directly to the master database.

Let FME do the work

FME can be used to sync your master data store with your open data repository. **Learn how to lessen your workload**.

Solutions Overview

WHETHER YOU'RE AN INDUSTRY PROFESSIONAL OR A HOBBYIST, THERE IS A SOLUTION FOR YOU.

Providers of open data are fortunate to have an abundance of top quality open data solutions readily available to them.

Each of the solutions presented here offers its own set of strengths for data publishers.

Since each solution is so ample, we're only going to touch on a few key points and encourage you to keep the conversation going! The world is always creating more options and each one offers something new.

Finally, be sure to do your own research and see which solution is right for you.

ArcGIS Open Data

Configure your own branded Open Data site on top of ArcGIS Server or ArcGIS Online.

ArcGIS Open Data will be of particular interest if you currently use Esri within your organization. **Esri has made it very easy to create and configure an open data site**, allowing you to focus on your strategy, policy and adoption rather than technical and operational concerns.

Data Publishing and Management:

Open data builds directly on top of your published ArcGIS services. The following data types are supported:

- Host ArcGIS Online feature services
- ArcGIS for Server feature services
- ArcGIS for Server map services
- Image services
- CSVs
- Socrata & CKAN hosted datasets (via the open-source Koop middleware)
- Other - Web maps, URLs, Word docs and PDF

Once you have decided which data to publish, you can tailor your site to look how you want with **configurable widgets** and a code view that allows further **customization**. You need to ensure all data in ArcGIS is clean and has good metadata.

Visualization

- View both the data and metadata in the browser.
- Data interaction is limited to sorting columns and filtering by a search query.
- Create simple histograms, line, donut or scatter charts to look for patterns without downloading the data.

Geospatial Features

All data is downloaded in WGS 84, and you can also download the data as KML, Shapefile or via the API (JSON, Geoservice, WMS).

If you don't want to download the data, you can **load any spatial dataset into the ArcGIS web map viewer and get an extremely rich set of tools to visualize and analyze the data**.

Clients:
- opendata.dc.gov
- imap.maryland.opendata.arcgis.com

Licensing	Delivery Model
Commercial	SaaS

CKAN

Open source data portal - providing tools to streamline publishing, sharing, finding and using data.

CKAN is a **leading open source data portal** with over 300 open source data management extensions. It is a powerful platform **best suited for large organizations,** as it is relatively complex to set up and maintain.

Flagship sites:
- data.gov
- data.gov.uk
- data.gc.ca

CKAN and FME Server: A Proven Pair

FME UC 2014: FME Powers CKAN Open Data Portal

0:00 / 26:48

The City of Surrey has partnered with FME Server to allow any CKAN dataset to be downloaded in any format and any projection. See how they did it.

Data Publishing and Management

Getting data into CKAN is a relatively simple process. You can upload data via custom spreadsheet importers or do it directly via the web interface. There is also a rich **JSON API** that you can integrate with FME, letting you load any dataset that FME supports into CKAN. You can also import data to CKAN from other services at regular intervals.

Once uploaded, there are tools to manage the permissions and edit the metadata.

Visualization

A **rich search experience** allows you to quickly find the data you want. For each dataset, you can then review the metadata and inspect the data in a tabular, graphical and mapping view.

Geospatial Features

If the data has a geographic component, **CKAN can plot the data on an interactive map** so users can view a sample of the dataset and interrogate individual records. Geospatial search is also supported with a user being able to filter and search for data based on a geographic location.

FME Server can be integrated with CKAN to extend format and projection support.

Licensing	Deployment Model
Open Source	Self-hosted, with SaaS offerings based on the CKAN technology.

DKAN

Complementary offering to CKAN. Open source Drupal plugin with a cloud-hosted version available.

Used by data.gov.uk, DKAN **integrates CKAN features into Drupal** and has a complete set of content management features. DKAN is **simpler to deploy** and maintain than CKAN. It's particularly suited to anyone using Drupal.

There is also a cloud version of DKAN that lowers the deployment barriers further.

Licensing	Deployment Model
Open Source	Self-hosted / SaaS

DataPress

WordPress and CKAN seamlessly integrated and hosted on the cloud.

DataPress integrates CKAN features into WordPress, the world's most popular content management system, to make **data publishing fast and simple**.

WordPress integration allows you to **write blog posts, design pages, and manage menus and content in a simple web interface**. You can use off-the-shelf WordPress themes and choose from thousands of extensions to ensure the website powering your open data site looks and works exactly how you want it to.

Cloud hosting means you can launch a CKAN-powered portal in minutes and rely on the high availability and durability that the cloud offers.

Licensing	Deployment Model
Commercial	SaaS

Socrata

A platform that turns data into a utility that can be discovered, consumed, visualized, analyzed, and shared.

Data Publishing and Management:

It's very easy to publish data to Socrata using a WebUI, desktop sync tool, or API. You can upload CSV, Excel, and TSV files natively and Socrata offers support for Shapefiles, KML, KMZ, and GeoJSON. There is also an FME Socrata Reader /Writer for uploading data. Other file types (such as PDF) can be uploaded for inclusion in a catalog but their contents cannot be searched. Though you can manage the metadata. With Socrata, **there is both a published and working copy of the dataset** as well as lots of tools around metadata management and workflow. This allows specifications to be created for your organization.

Open Source and Interoperability

All Socrata technologies are developed in the open, on the company's Github page, and the majority, including the core API server and all installable clients, are open source licensed and free to use or modify. Socrata also organizes large-scale open beta programs to solicit the input of governments around the world when developing new services.

Flagship sites:
- [NYC](#) | [Washington State](#)

Watch the Video: "Public Data as a Utility"

Visualization

Socrata boasts an efficient search that gives clear indication of what type of data has been found. Once inspecting a dataset there are a **rich set of tools that allow you to visually inspect both tabular and geospatial data**.

Unlike other platforms, **the metadata is not front and center; the focus is on the information within the data itself**. For tabular data you can set up filters and configure the data (i.e. select columns) in your preferred way before downloading. You **can also do charting on the data within the WebUI**, producing powerful visual representations of the data without having to use external tools. For users that prefer an external tool, all Socrata datasets include an API and an OData endpoint, which can be used to interact with the data using Excel, PowerBI, Tableau, and other **analytics tools**.

Geospatial Features

Socrata allows you to visualize and interrogate data from within the web browser. You can **overlay your own datasets** on top of the map and can save views that you have created for sharing at a later date. Socrata also **allows for a direct connection to an Esri catalog**.

Licensing	Deployment Model
Commercial	SaaS

OpenDataSoft

Focus on ease of use, automated API generation and interactive visualizations.

OpenDataSoft is a SaaS platform and provides a standard data portal as well as APIs that enable developers to integrate data into their applications.

They are proficient working with large datasets, as they leverage Elasticsearch, which ensures near real-time search and analysis.

Customers:
- City of Paris
- City of Brussels

Functionality:
- Data publishing and management via live dashboards
- Rich visualization
- APIs automatically generated from your data
- Geospatial format support

Licensing	Deployment Model
Commercial	SaaS

Junar

Cloud open data platform with focus on ease of use, powerful analysis and visualizations.

Junar is a SaaS platform and provides a standard data portal as well as APIs that enable developers to integrate data into their applications.

Customers:
- Government of Chile
- City of Sacramento
- City of Palo Alto

Functionality:
- Data publishing and management
- Data analysis

Licensing	Deployment Model
Commercial	SaaS

Free Hosting

Free and simple to use.

Ah, the magic free word. If you are looking for data visualization or analysis then look to the previous solutions, but this is a good fit if all you want is a simple file catalog service. **Free hosting with sites like DataHub.io, FTP, Google Drive and GitHub is a good place to start**.

The cloud file storage solutions can be used **to store and serve large volumes of data**. It is simple to upload the data, and a simple web interface can be built on top of the storage system to provide further context.

If collaboration with users is important, look at GitHub. Several people have successfully used it to host open data, and by uploading GeoJSON, you can even inspect the data on a map. In all cases, FME can be used to sync the storage services with the master database to ensure the databases are up to date.

Example:
- Github - State of Rioja
- FTP - Vancouver

Licensing	Deployment Model
Commercial Free	SaaS

Amazon Web Services (AWS)

Run your own powerful open data platform.

By leveraging the lower-level services of AWS (e.g. S3, EC2, RDS) and making use of FME Cloud as a data-mover, you can produce an extremely **fault-tolerant, scalable and powerful service that is easy to maintain and cost effective.**

Extensive format and projection extraction choices can be supported via FME Cloud.

Architecture

- Vector data: Stored in PostGIS/SQL Server Spatial RDS Database. FME can connect to almost any data store, so you can leave your data where it is.

- Raster/LiDAR data: Stored in AWS S3 with the footprint stored in the vector database for quick querying.

- FME Cloud is used to manage extraction requests.

Cost Analysis

Flexible, pay-as-you-go pricing on both AWS and FME Cloud makes it a great option if you are conscious of your spending but still want rich functionality.

Below is a comparison showing the costs of running the application on-premises vs. the cloud.

On Premises	Monthly	Yearly	Up-Front
DIS Server Space	$3,200		
FME Dekstop		$6,000	
FME Server		$12,000	
Dell Hardware			$100,000
SQL Server			
Windows Server (Software Assurance)		$595	
Windows Server Enterprise (Software Assurance)		$385	
SQL Server (Software Assurance)		$270	
SQL Server (Licenses)		$294	
MS Server Std Edition (License)		$1,662	
MS Windows Server (Licenses)		$4,490	
Symantec		$680	
Tape Backups (No server)			$2,000
Total Recurring Montly Payment	$3,200		
Total Recurring Yearly Payment	$26,376		
Total One Time Cost (Every 3 yrs)	$102,000		
Total Three Year Cost	**$296,328**		
True Monthly Cost (/36 months)	$8,231.33		

Intangiable Costs
Hardware Maintenance Time
DIS Process
Non-Scaleable

Cloud	Monthly	Yearly	Up-Front
FME Dekstop		$6,000	
FME Cloud		$14,400	
EC2 Instance (m3.large)	$395.81		
AWS Storage (EBS 780GB)	$77.11		
AWS Storage (S3 2TB)	$61.30		
AWS Storage (Glacier 4.2 TB)	$42.41		
Total Recurring Montly Payment	$576.63		
Total Recurring Yearly Payment	$20,400.00		
Total Three Year Cost	**$81,958.68**		
True Monthly Cost (/36 months)	$2,276.63		

> *Our rack space (real estate on our data center floor) costs $3,800 per month. Add to that the hardware costs, etc. and you start to see why moving to the cloud was a no brainer for us.*
> **- Anthony Davis, State of Arkansas**

Licensing	Deployment Model
Commercial	PaaS
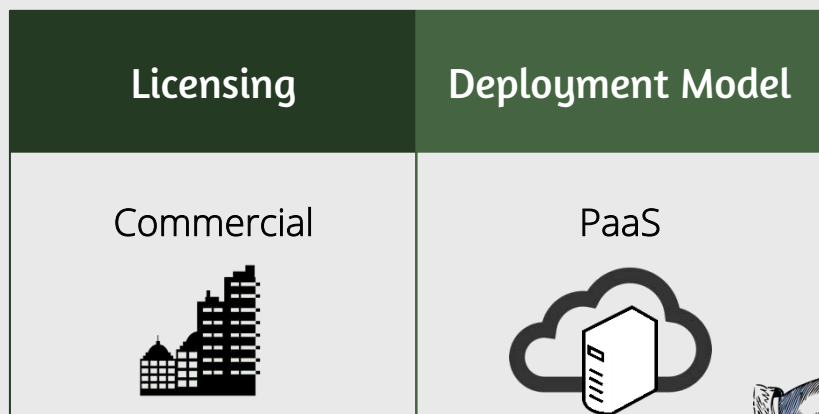
Data Migration Project: GeoStor

Goal: Create cloud-based Open Data catalog for the state datasets that is downloadable in multiple formats and projections.

Results:

- 300 vector datasets successfully migrated to PostGIS RDS
- 3TB of raster data migrated to AWS S3
- 4TB of historical raster data migrated to AWS Glacier

How it was done: Bulk one-time FME data migration workflows were created that loaded data from on-premises data stores to AWS services. These cloud databases and data stores then became the master data stores with updates applied directly to them.

A custom WordPress website was then created that allowed users to order data for a specific area in the format and projection they required. This was integrated with FME Cloud using the REST API, which handled all data extraction requests.

Because FME Cloud dynamically extracts from the master database, the data is always up to date. The full architecture is outlined in the adjacent diagram.
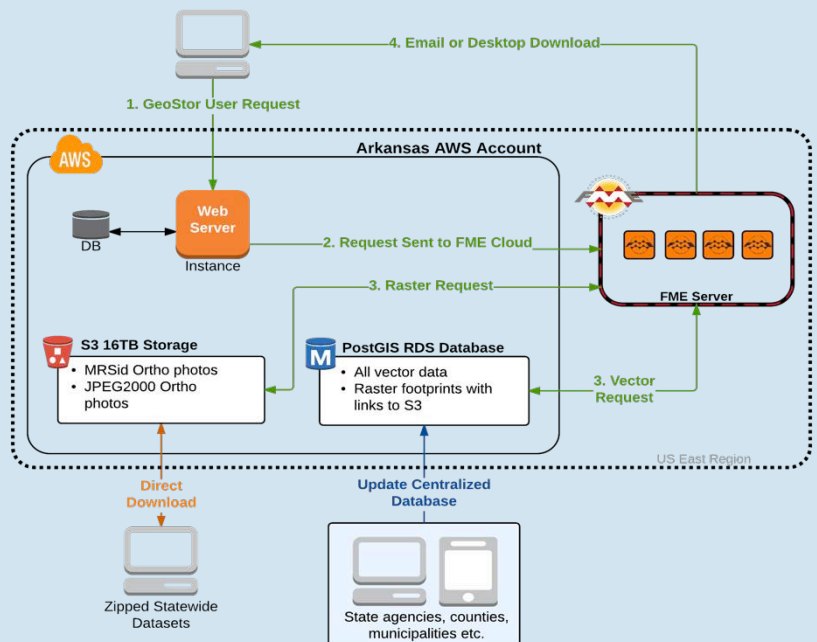
The Future of Open Data

Continuing towards greater global knowledge, interoperability and government transparency.

This is an exciting time for open data. With technology continuing to improve, and social media continuing to rise, users are able to explore and share their data at a new level. Engagement is easier than ever, as recipients are not limited to one format or server.

As a result, the public is starting to demand more information to be shared. While open data users can take advantage of this, even the citizens with no knowledge of open data can use it. It might be to research the crime rate before purchasing a new home, to find out where their tax dollars go, or to find out how much members of parliament make. The possibilities are endless, so long as we continue to share and update open data.

By keeping open data alive, new trends will start to develop on a global level. They've already started with the Open Private Sector and relief efforts.

Important trends moving forward:

Events like *hackathons* will boost in popularity among the open data community. In a hackathon, a group of programmers, developers, and designers work together to innovate new technology in a tight timeframe. New apps, video games, and operating systems are often the result.

Data sharing will become the norm at a national level, though there will be a push for globalization so we can compare cities and countries.

The Open Private Sector (a non-governmental suite of open data, open knowledge and open web) will emerge, driving a new era of transparent business.

Focus will shift to data quality and not quantity. It is easy to publish a dataset but much harder to keep it up to date or even provide real-time feeds.

Resources & Demos

ArcGIS
Demo

GeoStor (AWS)
Demo

DataPress
Demo

Socrata
Demo

Saxony
Open Data Portal

Sharing Open Data
on GitHub with FME

TED Talk
Smart Data NYC

Webinar
Open Data Portals

Want to know more about moving data?

View hundreds of articles and tutorials on our Knowledge Center, or explore our popular blog.

SAFE SOFTWARE™

Safe Software Inc. is the maker of FME and your global leader in spatial data transformation. From small businesses to large enterprises, we have an interoperability solution for you.

FIND YOUR DATA SOLUTION