STATE OF MARYLAND
DEPARTMENT OF INFORMATION TECHNOLOGY

LARRY HOGAN
Governor
BOYD RUTHERFORD
Lieutenant Governor

DAVID A. GARCIA
Secretary

To: Members Council on Open Data

From: Barney Krucoff & Josh Exler, Staff to Council

Date: April 10, 2015

RE: Material for April 14, Council meeting

Agenda/Discussion Topics:

- **Meet the new DoIT Secretary: David Garcia**
- **January 2015 Annual Report recommendations & status (Barney Krucoff).**
- **StateStat and its relation to Open Data (Kevin Conroy, StateStat Director)**
- **Automating StateStat Data Uploads to the Open Data Portal (Josh Exler)**
  - ***Advantages of Accessing StateStat Data through the Open Data Portal***
  - ***Brief Overview of Process to Copy Data from Agencies' Excel Spreadsheets to the Open Data Portal***
  - ***Automated Analysis of Agencies' Data***
  - ***Technical Considerations as Template Data are Added, Removed, Changed.***
  - ***Data which should Not be Public***
  - ***Conclusion***
- **New Data Visualization Tools on data.maryland.gov Coming April 22nd (Clare Zimmerman, Socrata)**
- **Data Freshness Dashboard (Josh Exler)**
- **Automating Large Dataset Publication (Josh Exler)**

- **Meet the new DoIT Secretary: David Garcia.** DoIT's new Secretary, David Garcia, will be in attendance. The Secretary has not been extensively briefed on the Council, so this meeting is opportunity for members to help familiarize the Secretary and through him the new Administration.

  Possible topics for discussion (this discussion may be moved to the end of the meeting):
  - Member's and Secretary Garcia's thoughts on open data?
  - Member's and Secretary Garcia's thoughts on what the Council should accomplish during the coming year and during the Hogan Administration longer term.

- **January 2015 Annual Report recommendations status (Barney Krucoff).** As per Maryland's Open Data Policy, SB644, the Council on Open Data is required to submit an annual report to the Legislature detailing Council topics of discussion to date and recommendations for actions items

within the following year. The Council submitted its first such report to the Legislature in January 2015, and made recommendations on the topics shown in the table below. Since the previous Open Data Council meeting (November 2014), some items have seen more progress than others. Status updates for the recommendations are also shown in the table below.

Recommendations from Open Data Council Annual Report – April 2015 Status Updates	
Recommendation	**Status Update**
• **SB0094/HB0353 –** The Council recommended reducing state and local governments' ability to charge for geographic data and eliminating the requirement that government entities enter into contracts with data users.	The Senate and House passed the legislation, sponsored by Senator Bill Ferguson and Delegate Bonnie, on February 24th and April 2nd, respectively. The Senate passed an enrolled version of SB94 on April 3rd. All votes were unanimous. The legislation is yet to be signed into law.
• **Retire executive orders 01.01.2012.04 and 01.01.2012.18.**	This recommendation has not yet been taken up by the new Administration.
• **Online open meetings calendar**	DoIT has contacted the Office of the Secretary of State about receiving a feed of open meetings. SOS has not yet taken up this initiative.
• **Statewide data inventory**	DoIT has used feedback from the previous Open Data Council meeting to design a preliminary form to log a dataset within the Statewide Data Inventory. DoIT is in the process of building an internal inventory database to house agencies' submitted datasets. DoIT plans to use the existing broader inventory of agencies' databases, performed by DoIT to document and implement cybersecurity solutions, as a starting point for the backend of the database for the Statewide Data Inventory.
• **DoIT should consolidate/synchronize http://data.maryland.gov & http://imap.maryland.gov**	DoIT will add to the data catalogue on data.maryland.gov external links to iMAP, one for each map. This will allow users to search data.maryland.gov as a consolidated source for all of Maryland's open data – geographic and non-geographic alike. Additional integration may be possible/desirable and DoIT is in discussions with Esri and Socrata.

Possible topics for discussion:

- Other 2015 legislation of interest to members?
- Any thoughts on 2016 legislation?
- Do any agencies have existing inventories or lists of their information resources, i.e., databases?

- What features would the Council like to see included if the Open Meetings Calendar were to be published in an open data format?

- **StateStat and its relation to Open Data (Kevin Conroy, StateStat Director).** Kevin Conroy, the new Director of StateStat, will explain how the new Administration's performance management initiatives and priorities will differ compared to those of the previous Administration.

  Possible topics for discussion:
  - What will be the main emphasis of StateStat under Governor Hogan?
  - How does Director Conroy see open data relating to StateStat's priorities?

- **Automating StateStat Data Uploads to the Open Data Portal (Josh Exler)** DoIT and StateStat have begun the process of copying the data currently reported to StateStat Excel spreadsheets, i.e., data templates, to datasets on the Open Data Portal, data.maryland.gov. The goal is provide automated analysis to both agencies and StateStat analysts, without requiring any additional work from agencies nor changes to existing workflows. Use of the open data portal has the additional benefit of making public access and interpretation of the data easier.

  - ***Advantages of Accessing StateStat Data through the Open Data Portal.*** Consolidating StateStat data to data.maryland.gov will create a central repository for StateStat data and facilitate public user – as well as agency – access to the data. After testing several strategies for automating the data upload process using data templates from the State Highway Administration (SHA) last year, StateStat and DoIT are proceeding with a finalized set of scripts to automate an Extract, Transform, and Load (ETL) process. Department of Health and Mental Hygiene (DHMH) data are the first being uploaded.

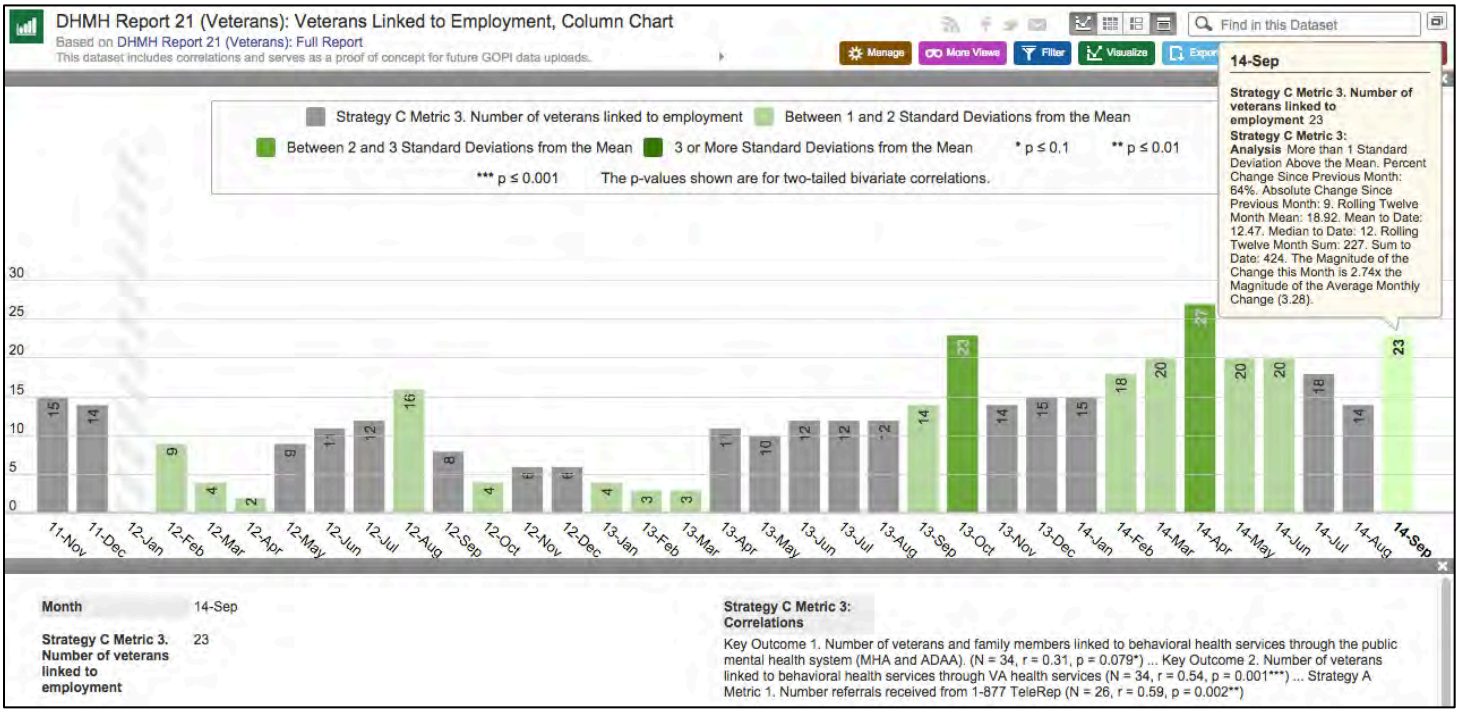
  - Agencies' Excel spreadsheets, i.e., data templates, submitted regularly to StateStat, vary widely in data formatting and presentation. The primary means by which the public had access to these data previously was through StateStat posting the Excel spreadsheets to its website for download. These data were locked, however, and therefore geared towards viewing the data instead of working with the data. **The Open Data Portal provides a standardized and accessible format for all agencies' StateStat data moving forward.** Graphs of agency data will live in static locations within the Open Data Portal, automatically updated when the agencies update their Excel spreadsheets, thus eliminating a disconnect between agencies' analysis of their own data and StateStat's analysis.

- ***Brief Overview of Process to Copy Data from Agencies' Excel Spreadsheets to the Open Data Portal.*** The table below shows a broad overview of how DoIT copies agencies' template data to data.maryland.gov. Many more details, including what agencies should keep in mind moving forward to make sure their data can be accessed by DoIT, are included in the next sections.

Scripts to Copy StateStat Data to the Open Data Portal

Step 1: Extract. DoIT has used the scripting language Visual Basic to automate the process of copying all data within DHMH's data templates. Whenever an agency submits new data templates to StateStat, DoIT will run this script to copy the new data.
Step 2: Transform. The copied data are structured and formatted in a Socrata-readable format, e.g., by removing blank columns and rows from agencies' spreadsheets. At this stage, the Visual Basic script performs an automated data analysis on all of the agency data (see next section).
Step 3: Load. The transformed data are uploaded to Socrata using the software FME. The ability for FME to write to Socrata datasets is powered by the Socrata Publisher API.

- ***Automated Analysis of Agencies' Data.*** The graph below shows an example of the automatic analysis performed as part of DoIT's Extract, Transform, and Load (ETL) scripts. Agencies and StateStat will have access to the same analysis. The table below the graph provides details on what analyses are included. A live version of the graph can be found at https://data.maryland.gov/Health-and-Human-Services/DHMH-Report-21-Veterans-Veterans-Linked-to-Employm/4ymr-b22x. Accessing the graph through this link is preferable to only viewing the screenshot below, since the link allows the user to see the automatically generated analysis on each month's data and shows the interactive capabilities. Hover over each column to see the automatically generated analysis for that month. Alternatively, use the buttons at the bottom of the page to page through the 35 months' worth of analysis.



Automated Analysis of StateStat Data – Included in Datasets on Open Data Portal
This analysis will be generated for every dataset containing StateStat data which is uploaded to data.maryland.gov. Each data series has a new analysis automatically generated every time new data templates are submitted to StateStat.

• Automatic detection of non-numeric data (this will help detect and fix fat-finger errors)
• Absolute Change (compared to previous data point, e.g., how March's new data compare to February's data)
• Relative Change (the same change, but measured as the relative difference and reported as a percentage)
• Average Change to Date (e.g., average monthly change to date)
• How the Current Change Compares to the Average Change (this will help analysts detect unusually large increases or decreases in a data series)
• Standard Deviations from the Mean – Sometimes a large monthly increase or decrease happens only because a data series has high variance. Automatically calculating standard deviations from the mean will show whether increases are due to expected variance – or rather, whether increases and decreases are driven by external causal factors.
• Mean to Date
• Median to Date
• Rolling Mean (the default period used is a 12 month rolling mean for monthly data and a 4 year rolling mean for annual data)
• Sum to Date (not calculated for percentages or rates)
• Rolling Sum (not calculated for percentages or rates)
• Bivariate Correlations – Each pair of data series within a dataset[1] is tested for correlation. The coefficient of correlation (R squared) is reported, as is the p-value and sample size (N). The latter value shows the number of data pairs included in the test. By default, bidirectionality of causal relationships is assumed[2], so two-tailed tests are used to test for statistical significance, with any p-values less than or **equal** to 0.1 flagged.[3]

The automatic analysis is a new toolkit for StateStat and agencies to detect, flag, and dive deeper into data trends without resorting to manual arithmetic calculations and statistical analyses performed on hundreds, or thousands, of data series per month per agency. By devoting less time to manual analysis, StateStat should be able to focus more on policy and process.

Possible topics for discussion:

- Does the proposed method for automating the existing Excel based work processes seem workable?
- Will the automated analyses be helpful to agencies?
- Are there any additional tests, analyses, or calculations that StateStat or agencies regularly use when looking at their data, which should also be included as part of the automatic analysis?

[1] Each dataset on data.maryland.gov roughly corresponds to a tab of an agency Excel spreadsheet.
[2] The scripts can be tweaked to test for one-tailed significance tests for any pairs of data series where a direction of a causal relationship is known.
[3] 0.1 is a liberal upper bound for statistical significance with a two-tailed test. Significance at or below the 0.01 level, as well as at or below the 0.001 level, is also flagged.

Do agencies have questions about any of the components of the automatic analysis, e.g., how a calculation or analysis is performed, or what utility the component has for performance management?

- ***Technical Considerations as Template Data are Added, Removed, Changed.*** In order to take advantage of agencies' existing workflows to populate their data templates, DoIT's scripts to copy data from Excel spreadsheets to the Open Data Portal. In fact, the more agencies maintain their previous processes for populating their StateStat data to their Excel spreadsheets, the easier it is for DoIT to capture the data, since DoIT's scripts are based off of current data templates' layouts. With that in mind, several considerations will be needed when agencies want to add or remove any of the data series reported to their StateStat templates:

  - ***Adding Data:*** While DoIT would prefer that agencies add new data series to the end of a spreadsheet tab, i.e., as a new last row or last column, in most circumstances it will be feasible for DoIT to maintain its scripts if the agencies need to insert a new row or a new column in the middle of a tab.
  - ***Removing Data:*** Instead of deleting a row or a column outright, or "blacking out" data, i.e., shading cells black, simply hide the row or column, and DoIT will hide the corresponding data in the Open Data Portal.
  - ***Modifying Data:*** If an agency needs to change the name of a data series, it should inform DoIT so that DoIT can change the name of the corresponding data series in the Open Data Portal.

DoIT will be meeting with each agency in advance of copying their data to the Open Data Portal, to go over these requirements in detail. The key fact to keep in mind is that **DoIT's scripts point to specific cells within Excel spreadsheets, so if agencies change the location of data within Excel without informing DoIT, the copying of data to the Open Data Portal will not complete successfully.** Unreadable data will be flagged automatically and returned to the agency for cleanup. DoIT and StateStat will work closely with the agencies to make sure their Excel spreadsheets' data remain readable by DoIT's scripts.

Possible topics for discussion:

- Have any agencies added or removed any data from their data templates since November 2014?
- Would agencies be open to adding their StateStat data directly to the Open Data Portal, eliminating the need for Excel data templates entirely?

- ***Data which should not be Public.*** Data which the agencies determine to pose a security risk or contain Personally Identifying Information (PII) will never be made public on the Open Data Portal. This will be accomplished by hiding these data series from public views on data.maryland.gov (data series within a Portal dataset can be hidden). Agencies and StateStat will still be able to see graphs and analysis of the sensitive data, but the public will not. The Council should note that the Open Data Act specifies exactly five reasons that State data should not be made public in an open data format. Data meeting one or more of these criteria will be made private on data.maryland.gov.

Possible topics for discussion:

- Show of hands: How many agencies have regularly submitted PII or other types of sensitive data to StateStat?
- Show of hands: How many agencies have irregularly submitted PII or other types of sensitive data to StateStat?
- Do agencies have concerns about the security of sensitive data contained in private pages on data.maryland.gov?

- **New Data Visualization Tools on data.maryland.gov Coming April 22nd (Clare Zimmerman, Socrata).** On April 22nd, the Open Data Portal will gain a new way to visualize and interact with data which Socrata is calling "Data Lens." This will roll out to all Socrata-powered sites on that date. A preview of a Data Lens, created using an existing data.maryland.gov dataset, can be seen below. A live preview link is not yet available. Eventually, these Data Lens pages will become the primary landing pages for datasets. The screenshot below only shows part of a Data Lens; underneath more sections are included, including the data table for the backing dataset. Which sections, or which "cards", are included is determined automatically by Socrata, analyzing a dataset's structure and data types.

  - Clicking on areas of a map, or columns in a graph, automatically filters data on the whole page, thereby giving each Data Lens page dashboard functionality. The automatically generated Data Lenses will act as automatically-generated dashboards to visualize and interact with every dataset on data.maryland.gov.
  - Clare Zimmerman will show several live examples of Data Lens created using data.maryland.gov datasets and answer questions about their current features, as well as what future capabilities are planned. Socrata has scheduled a webinar for April 16th from 1 pm to 2 pm EST to allow Socrata users to learn more about Data Lens.

Possible topics for discussion:

- What datasets on data.maryland.gov would work best for visualization using Data Lenses?
- How do uses differ for geographic versus non-geographic data?
- When will Data Lens become the default landing page for data.maryland.gov datasets?
- Are any Council members interested in attending the April 16th webinar?
- Are additional training dates needed?

- **Data Freshness Dashboard (Josh Exler).** DoIT has used the Socrata Open Data API (SODA) to develop a "Data Freshness Dashboard" for data.maryland.gov. This dashboard is not yet live, but today's meeting includes a demo of the dashboard so that agencies can prepare for its public launch and update their datasets if need be. Once the dashboard is live, it will be prominently linked to on data.maryland.gov so that internal and external users alike can see whether datasets are up to date.
  - The data freshness dashboard evaluates whether each dataset on data.maryland.gov has been updated recently enough. Most open data portals' Achilles' heels are keeping their content up to date. The Open Data Portal is no exception, with many datasets not modified since an initial data dump in 2012 or 2013. Data.maryland.gov has a custom metadata field called "Update Frequency" which must be filled out upon dataset upload, indicating whether datasets are going to be updated weekly, monthly, quarterly, semiannually, or annually ("As needed" and "Unknown" are two additional options). The Data Freshness Dashboard compares this metadata to the date of last dataset update, to determine whether the dataset has been updated recently enough, e.g., within the previous 31 days for datasets which are supposed to be updated monthly. A screenshot of the dashboard is shown below and a demo will be provided at today's meeting. The dashboard includes a search feature to show all datasets from a specific account or state agency.

- The Council should note that names of individual data providers will not be included in the public version of the dashboard, only the internal version.

Maryland Open Data Portal: Data Freshness Dashboard

Show All entries
Search:

Dataset Name	Link	Owner	User who Made Last Update	Update Frequency	Date of Most Recent Data Change	Updated Recently Enough?	Number of Rows
Maryland Port Administration General Cargo	https://data.maryland.gov/d/2ir4-626w	Dominic Scurti	https://data.maryland.gov/profile/nqr5-s4py	Monthly	Thu Mar 26 2015 12:54:26 GMT-0400 (EDT), 13 days ago	Yes	206
BayStat Solutions Reporting - Maryland Dept. of Natural Resources	https://data.maryland.gov/d/4zqs-i2t2	ESRGC	https://data.maryland.gov/profile/xgdh-qy5z	Annually	Fri Mar 27 2015 14:02:34 GMT-0400 (EDT), 12 days ago	Yes	66
BayStat Solutions Reporting - Maryland Dept. of Natural Resources	https://data.maryland.gov/d/4zqs-i2t2	ESRGC	https://data.maryland.gov/profile/xgdh-qy5z	Annually	Fri Mar 27 2015 14:02:34 GMT-0400 (EDT), 12 days ago	Yes	66
BayStat Solutions Reporting - Maryland Dept. of the Environment	https://data.maryland.gov/d/ab68-n7ja	ESRGC	https://data.maryland.gov/profile/6wh5-kegk	Annually	Wed Apr 01 2015 16:29:12 GMT-0400 (EDT), 7 days ago	Yes	44
BayStat Solutions Reporting - Maryland Dept. of Agriculture	https://data.maryland.gov/d/tsya-25ee	joshuaexler	https://data.maryland.gov/profile/hcsr-zg76	Annually	Mon Mar 23 2015 11:38:31 GMT-0400 (EDT), 16 days ago	Yes	55

Information Included in Data Freshness Dashboard
• **Dataset Name**
• **Link**
• **Owner –** A dataset's default owner is the user who uploaded the dataset. Dataset ownership can be transferred to another user after the initial upload. The owner should be whoever is directly responsible for maintaining the dataset.
• **User who Made Last Update –** This is sometimes the dataset owner, but in other cases, agencies assign multiple users to make updates to a dataset.
• **Update Frequency –** DoIT is planning to add "Static Data" as an option for data which is not planned to be updated after the initial upload.
• **Date of Most Recent Data Change**
• **Updated Recently Enough? –** This information shows as "Yes", "No", or "N/A" for datasets where the update frequency is logged as "As needed."
• **Number of Rows – Arguably the least important part of the dashboard**
• **(Planned but not yet Included) Tags/Keywords**

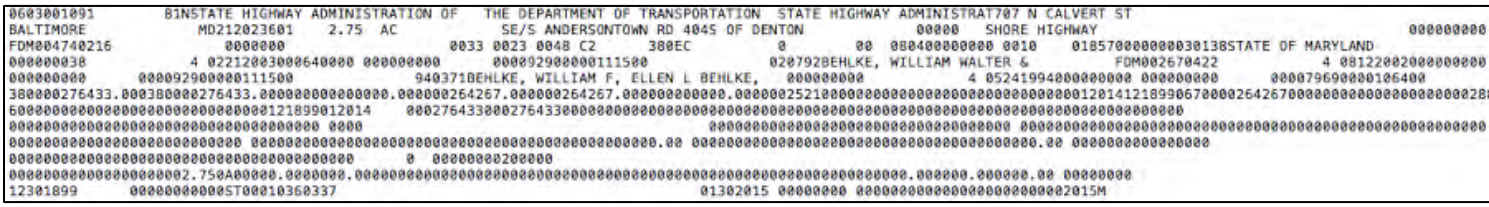
Possible topics for discussion:

- Is there any other metadata which should be included in the dashboard?

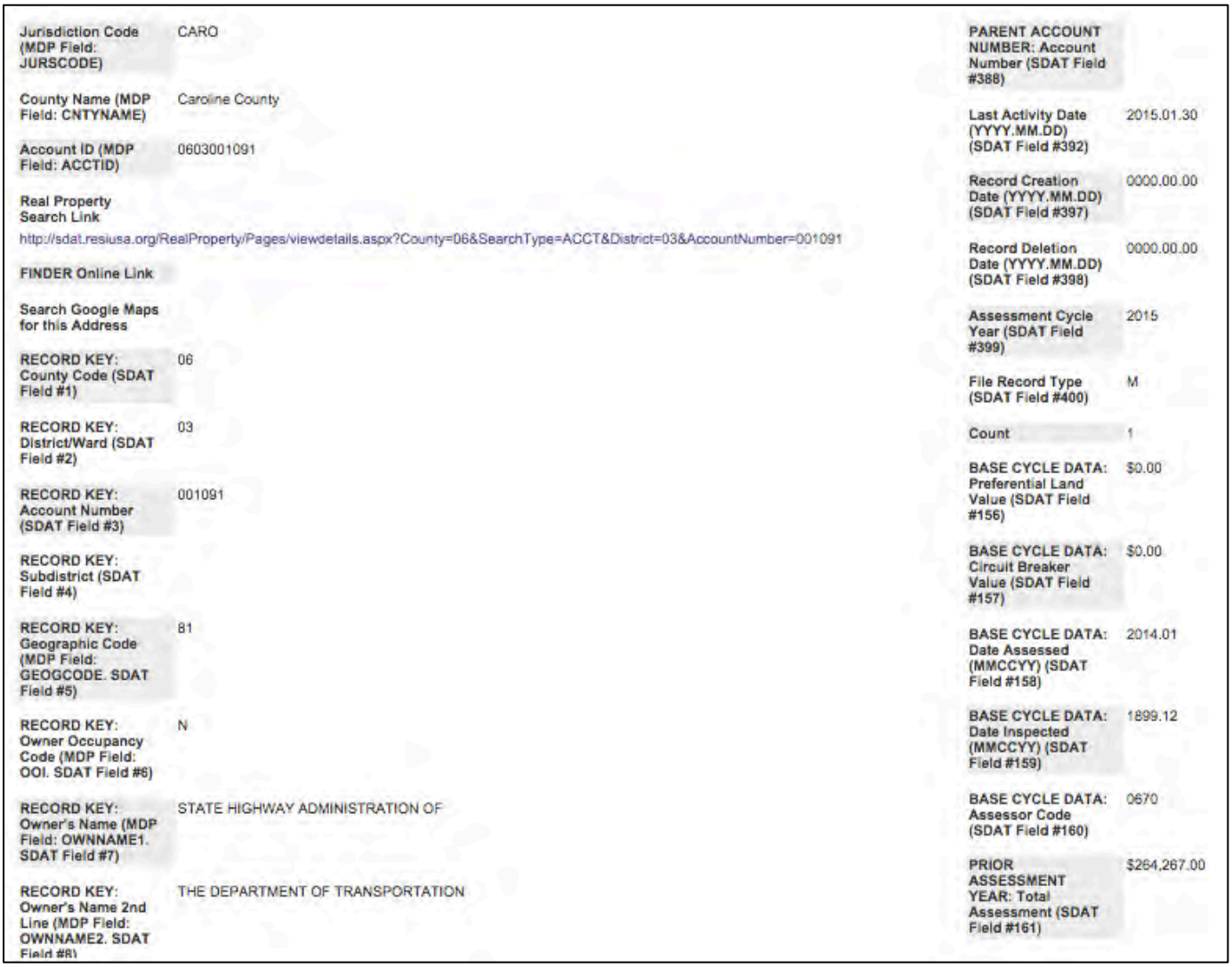
- What are agencies' current plans and approaches to making sure their data are updated regularly and do not fall out of date?

- Does the Council have any questions about any of the content in the dashboard or how any of this information is retrieved or calculated?

- **Automating Large Dataset Publication.** DoIT has begun using the software FME (Feature Manipulation Engine) to automate Extract, Transform, and Load functions from state agency databases to data.maryland.gov. As an initial project, DoIT has begun receiving monthly data extracts from the State Department of Assessments and Taxation (SDAT) on statewide property records and assessments. DoIT has automated the transformation of each property's record – received from SDAT as a string of numbers and letters (see first screenshot below, which represents a single property) – into records in a data.maryland.gov dataset, with several hundred fields included. The information in this Socrata database will be updated when DoIT receives updated data from SDAT at the beginning of each month. Once the dataset is public, anyone in Maryland will be able to access SDAT's entire collection of property records in a consolidated dataset. Whereas all existing products to access these data only show one property at a time, the data table on data.maryland.gov will show all of the millions of property records in the state in one location, with the ability to graph, filter, map, and export any of the data – all updated monthly. DoIT is currently working with MDP to finalize the dataset structure and formatting before passing a final version along to SDAT for approval. Once all three agencies have signed off, property records will start to be made public, county by county.
  - A similar approach can be used by DoIT to automate the publication of large datasets from other agencies. DoIT currently receives data from SDAT through an FTP server, which at this time is the ideal means to transmit the data. Eventually DoIT should be able to use FME to build direct connections to agencies' internal databases without the need for preparatory data extractions to FTP servers or external locations.



Screenshot of an SDAT property assessment record



Jurisdiction Code (MDP Field: JURSCODE)	CARO	PARENT ACCOUNT NUMBER: Account Number (SDAT Field #388)	
County Name (MDP Field: CNTYNAME)	Caroline County	Last Activity Date (YYYY.MM.DD) (SDAT Field #392)	2015.01.30
Account ID (MDP Field: ACCTID)	0603001091	Record Creation Date (YYYY.MM.DD) (SDAT Field #397)	0000.00.00
Real Property Search Link	http://sdat.resiusa.org/RealProperty/Pages/viewdetails.aspx?County=06&SearchType=ACCT&District=03&AccountNumber=001091	Record Deletion Date (YYYY.MM.DD) (SDAT Field #398)	0000.00.00
FINDER Online Link		Assessment Cycle Year (SDAT Field #399)	2015
Search Google Maps for this Address		File Record Type (SDAT Field #400)	M
RECORD KEY: County Code (SDAT Field #1)	06	Count	1
RECORD KEY: District/Ward (SDAT Field #2)	03	BASE CYCLE DATA: Preferential Land Value (SDAT Field #156)	$0.00
RECORD KEY: Account Number (SDAT Field #3)	001091	BASE CYCLE DATA: Circuit Breaker Value (SDAT Field #157)	$0.00
RECORD KEY: Subdistrict (SDAT Field #4)		BASE CYCLE DATA: Date Assessed (MMCCYY) (SDAT Field #158)	2014.01
RECORD KEY: Geographic Code (MDP Field: GEOGCODE. SDAT Field #5)	81	BASE CYCLE DATA: Date Inspected (MMCCYY) (SDAT Field #159)	1899.12
RECORD KEY: Owner Occupancy Code (MDP Field: OOI. SDAT Field #6)	N	BASE CYCLE DATA: Assessor Code (SDAT Field #160)	0670
RECORD KEY: Owner's Name (MDP Field: OWNNAME1. SDAT Field #7)	STATE HIGHWAY ADMINISTRATION OF	PRIOR ASSESSMENT YEAR: Total Assessment (SDAT Field #161)	$264,267.00
RECORD KEY: Owner's Name 2nd Line (MDP Field: OWNNAME2. SDAT Field #8)	THE DEPARTMENT OF TRANSPORTATION		

Preview of how the same record will appear on the Open Data Portal. Note that several hundred additional data fields are cut off the bottom of the screenshot.

Possible topics for discussion:

- DoIT believes we are setting an important precedent by combining data from SDAT and MDP into a single dataset. Can other agencies think of other opportunities?
- Does the Council have questions on the means by which DoIT has modified SDAT's data?
- Can DoIT assist any additional agencies with the publication of their data to the Open Data Portal in similar projects?
- Could a similar process help agencies automate existing workflows for data already being published to the Open Data Portal?