

## Draft Data Aggregation Project Plan

Project Name: Data Aggregation 'Proof of Concept'

Publication Date: November 5, 2015

### Project Description:

#### What is the need or opportunity?

A number of workflows exist or are proposed which involve locally produced datasets being aggregated and published by a regional or state agency. Data produced at the local level and brought together regionally or statewide has been identified as critical for many GIS users, agencies and cross jurisdictional initiatives (e.g. E9-1-1). Datasets identified as critical to these groups and efforts include address points, street centerlines, parcels and park boundaries. This list of critical datasets is only expected to grow as technologies, standards and policies continue to be developed. Maintaining these aggregated resources will require an automated, consistent and streamlined aggregation strategy.

Due to the advancement in GIS data sharing technologies, the development of the GeoCommons website (a data portal for all MN GIS users), and the strong collaborative relationship between local, regional and state partners, an opportunity currently exists to develop a simple, automated and streamlined aggregation strategy. The strategy will include the ability to aggregate data regionally or on a statewide scale. It will be important to develop a strategy that does not require different methods of supplying aggregated resources or submitting resources to an aggregated dataset depending on the published aggregator (e.g. MetroGIS vs MetCouncil vs. MnGeo, etc.). Having a common approach will encourage participation from data producers thereby expediting the development of aggregated datasets and putting the resources in the hands of users.

#### Overall data aggregation goal

To develop technologies, workflows and inter-agency agreement(s) to facilitate the efficient and continuous (sustainable) standardization, aggregation, publication and availability of geospatial data.

## **Business Need or Drivers**

### ***How will this benefit stakeholders?***

There is a business need to implement a simple, consistent aggregation workflow for all stakeholders including:

1. Data Producers – Staff resources are needed to support current manual aggregation workflows which will soon become unsustainable as the number of aggregated datasets requested by end users continues to grow. Producers will be more willing to participate if the methods and tools used to contribute data to an aggregated resource are simple and consistent. This will promote broad participation among more data producing agencies throughout the state.
2. Data Consumers – Developing a consistent workflow for data producers will expedite the process of gathering data from a local source bringing it together and publishing in a consistent format. This will provide all users with more current information and extended data coverage as more producers are willing to participate.

Other benefits include:

- The authoritative data being readily available in as near to ‘real-time’ as possible/feasible to all users who desire it;
- Better use of staff time and agency resources in the data producer community as far as the ease of meeting demand for data
- An enduring example of how governments at all levels can collaborate and work together to best serve the public

## **Scope of Project**

The overall scope of the project is to create a proof of concept workflow using the address points dataset supplied by two data producers (Carver and Dakota Counties) and publishing the dataset as an aggregated resource on the Minnesota Geospatial Commons. One primary goal of the proof of concept is to demonstrate that the workflow can be scaled up to include additional datasets, producers or aggregators. This project begins with the assumption that an address point dataset has been generated by a data producer and that the producer is ready to contribute their data to an aggregated dataset.

Project Scope:

1. Address points formatted to the MetroGIS Address Point Specification
2. Data will be provided by two county data producers.
3. The published aggregator will be MetroGIS.
4. Data updated will be submitted and aggregated daily (or as near to daily as possible).

5. Tools will be developed to validate data producers' datasets two times within the workflow. The first validation will occur before data is submitted. The second validation will occur after the data has been submitted but before aggregation.
6. Data producers will be provided a tool to validate address point schemas before the data has been submitted. Error messaging will be incorporated.
7. Validation tools will check data schemas as defined by the MetroGIS address points specification.
8. Validation tools will check domain compliance at the record level.
9. Validation tools will check for other issues such as missing records, null fields, null geometry and corrupt geometry.
10. Data producers will be provided a tool to format data into GeoCommons specifications and to submit data to a Geospatial Commons staging areas.
11. Data producers will receive automated messages if submitted resources do not meet schema or GeoCommons specifications.
12. Data producers will receive an automated message or be provided an alternative feedback method so it is known that a successful aggregation has occurred.
13. Once data has been submitted by the data producers, GeoCommons process will aggregate and publish the dataset using existing GeoCommons infrastructure.
14. Data will be published as a shapefile, file geodatabase, web service (format of which yet To Be Determined) or other format determined by the project team.
15. Metadata and supporting documentation will be designed to supported daily aggregation goals.
16. MetroGIS and MnGeo will lead the effort to communicate project goals, progress potential for scaling up process to local, regional and state partners.

This proof of concept project will not include the following:

1. Working with data producers to develop internal processes to prepare data to meet the MetroGIS address points specifications
2. Using other datasets other than address points
3. Using more than two participating data producers at this time

## **Risks**

*Describe areas of risk associated with the scope, roles, timeline or other aspects of the project*

1. Lack of resources to transform data into a standardized format within data producing counties;
2. Firewall rules and constraints;
3. Other competing or fractured aggregation strategies;
4. GIS Staff (analysts, specialists) have many other duties and responsibilities
5. Having to convert data from other/adjointing/adjacent/concurrent jurisdictions in different formats for agency use;
6. Updating metadata is often a manual process and can be time consuming (completeness, positional accuracy, etc.)

## Deliverables

1. Project plan document
2. Technical workflow diagram
3. Tools for data producers for validating data schemas, domains and other data specifications
4. GeoCommons processes to support data submission, aggregation and publishing
5. Metadata workflow documentation
6. Resource package to be used by future data producers to expedite data producer participation

## Project Participants and Roles

Name and Department	Roles	Project Responsibilities
<b>Hal Watson (DNR)</b>	Technical Lead	<i>Lead development of GeoCommons validation, aggregation and messaging scripts;</i>
<b>Brent Lund (MnGeo)</b>	MnGeo Liaison	<i>Ensure MnGeo resources are available to support workflow; Communicate progress to MnGeo staff; Provide feedback from MnGeo staff regarding alignment with statewide aggregation goals;</i>
<b>Jon Hoekenga (Met Council)</b>	Project Coordinator	<i>Develop validation scripts for data producers; Develop modified metadata workflow; Track project progress;</i>
<b>Zeb Thomas (DNR)</b>	Technical Resource	<i>Assist technical lead on the development of GeoCommons validation, aggregation and messaging scripts;</i>
<b>Joe Sapletal (Dakota County)</b>	Data Producer	<i>Participate in workflow design Test validation tools</i>
<b>Pete Henschel &amp; Chad Riley (Carver County)</b>	Data Producer	<i>Participate in workflow design Test validation tools</i>
<b>Geoff Maas (MetroGIS)</b>	Communication	<i>Communicate project process to stakeholders</i>

#	Task Name	Lead Agency	Status
---	-----------	-------------	--------

## Work breakdown structure

The work breakdown structure contains the basic planning, execution, and quality steps needed to complete the project. Estimates should be made for how long each task will take to complete as well as when it will start and finish.

#	Task Name	Task Lead	Estimated Hours of Work	Start	Finish
1	Document proposed workflow	Jon H	8		
2	Develop tool for data producers to validate schema and records	Jon H	24		
3	Review metadata and metadata update process to accommodate automated workflow	Jon H	16		
4	Test schema and record validation tool	Joe Sapletal Chad Riley	16		
5	Provide data producers with tools and other resources to prepare and validate datasets against GeoCommons specifications	Zeb Thomas	24		
6	Develop GeoCommons validation, messaging and aggregation processes	Zeb Thomas	80		
7	Ensure aggregated data is published daily on the GeoCommons website	Zeb Thomas	20		
8	Develop resource package for additional data producers	Hal Watson	40		
9	Communicate project progress with other MetroGIS and relevant stakeholders	Geoff Maas Brent Lund	TBD		

**Estimated Start Date:** (To Be Determined)

**Estimated Completion Date:** (To Be Determined)

On the following pages are the **Draft Detailed Work Breakdown Tasks**

<b>1</b>	<b>Document proposed workflow</b>		
1.1	<ul style="list-style-type: none"> <li>Share draft project plan with team, assemble comments make changes as necessary</li> </ul>	Metropolitan Council, MetroGIS	Complete
1.2	<ul style="list-style-type: none"> <li>Finalize project plan and workflow diagram (share with interested parties)</li> </ul>	Metropolitan Council, MetroGIS	Complete
1.3	<ul style="list-style-type: none"> <li>Complete detailed project task list</li> </ul>	Metropolitan Council	Complete
1.4	<ul style="list-style-type: none"> <li>Email detailed proposed project schedule to team members, schedule team meeting if necessary</li> </ul>	Metropolitan Council, MetroGIS	Complete
<b>2</b>	<b>Develop tool for data producers to validate schema and records</b>		
2.1	<ul style="list-style-type: none"> <li>Make final updates to address points validation tool (most of this coding already existing)</li> </ul>	Metropolitan Council	Complete
<b>3</b>	<b>Review metadata and metadata update process to accommodate automated workflow</b>		
3.1	<ul style="list-style-type: none"> <li>Prepare metadata recommendations documentation</li> </ul>	Metropolitan Council	Complete
3.2	<ul style="list-style-type: none"> <li><b>Team Meeting (Remote):</b> Metadata review, finalized changes</li> </ul>	Metropolitan Council	Complete
3.3	<ul style="list-style-type: none"> <li>Finalized metadata documentation and workflow</li> </ul>	Metropolitan Council	Complete
<b>4</b>	<b>Test schema and record validation tool</b>		
4.1	<ul style="list-style-type: none"> <li>Council to send ArcMap validation tool to data producers for testing</li> </ul>	Metropolitan Council	Complete
4.2	<ul style="list-style-type: none"> <li>Data producers test validation tool</li> </ul>	Data Producers (Counties)	Complete
4.3	<ul style="list-style-type: none"> <li>Make changes to tool based on data producers review</li> </ul>	Metropolitan Council	Complete
4.4	<ul style="list-style-type: none"> <li>Work with data producers as need to incorporate coding into automated updated processes if exist</li> </ul>	Metropolitan Council	Complete
<b>5</b>	<b>Provide data producers with tools and other resources to prepare and validate datasets against GeoCommons specifications</b>		
5.1	<ul style="list-style-type: none"> <li>DNR to provide data producers GDRS validation tools</li> </ul>	DNR	Complete
5.2	<ul style="list-style-type: none"> <li>MnGeo to provide data producers access and instructions for managing resources</li> </ul>	MnGeo	Complete
<b>6</b>	<b>Develop GeoCommons validation, messaging and aggregation processes</b>		
6.1	<ul style="list-style-type: none"> <li>Share Council validation coding with DNR for backend schema, and domain validation</li> </ul>	Metropolitan Council	Complete
6.2	<ul style="list-style-type: none"> <li>DNR to create incorporate schema validation code with messages code to complete workflow</li> </ul>	DNR	In Progress
6.3	<ul style="list-style-type: none"> <li>Data producers to modify automated scripting to incorporate daily ftp posting to GeoCommons</li> </ul>	DNR	In Progress
<b>7</b>	<b>Ensure aggregated data is published daily on the GeoCommons website</b>		
7.1	<ul style="list-style-type: none"> <li>Data producers to post resources for testing</li> </ul>	Data Producers (Counties)	Complete
<b>8</b>	<b>Develop resource package for additional data producers</b>		In Progress
8.1	<ul style="list-style-type: none"> <li><b>Team Meeting (Remote):</b> Get all feedback from data producers and team members to better inform DNR and Council for preparing data package for data producers beyond this project</li> </ul>	Metropolitan Council	
8.2	<ul style="list-style-type: none"> <li>Council and DNR to prepare package of tools, best practices and other instructions needed to quickly implement workflow</li> </ul>	Metropolitan Council, DNR, MetroGIS	In Progress
<b>9</b>	<b>Communicate project progress with other MetroGIS an relevant stakeholders</b>		
9.1	<ul style="list-style-type: none"> <li>Ongoing as needed</li> </ul>	MetroGIS & MnGeo	In Progress