Objective

The purpose of this document is to provide an understanding and background of what is considered a duplicate entity within the Ag Industry Identification System (AGIIS) and what processes and resources are in place to prevent duplicates.

Background

The purpose of AGIIS is to provide unique ID's for the facilitation of eBusiness interactions (i.e. both partners are confident they are using the correct and same id for a given entity). AgGateway's vision is for AGIIS to be the industry benchmark for accurate and trusted information. In order to accomplish this vision, the organization understands that you need to have sound processes in place to promote data quality and integrity. One of the key items that need to be in place is a duplicate prevention strategy. AgGateway has developed a multi-pronged approach to prevent duplicates. In this document we will provide some background on how we developed our duplicate prevention strategy and walk you through the processes that are in place to prevent and eliminate duplicates.

AgGateway has dedicated much time to understanding, preventing and eliminating duplicates within the AGIIS Directory. In 2011 AgGateway formed a task force to comprehensively review the issues of duplicate entities, both real and perceived. In September of 2011 the AGIIS Directory Oversight Committee approved 13 recommendations that would further enhance the AGIIS Duplicate Prevention process. These recommendations range from providing a definition of a duplicate entity to recommendations on how to report duplicates as well as enhancements to the current duplicate prevention processes.

What is a Duplicate?

The AGIIS Directory is a subscriber driven directory, which means the source of the data comes from the users of the database. As this is the case, it is critical that participants provide quality data and continue to provide updates and synchronize with AGIIS and their trading partners. If these activities are being performed by subscribers, there is a better chance that duplicates will not be introduced to the database. AgGateway recognizes that there will always be some level of duplicates that exist in the database but the goal is to keep those duplicate numbers to a minimum.

The first step is to understand what constitutes a duplicate. AgGateway has developed clear and concise definitions of an entity that takes into account the needs of the various industry segments within AgGateway as well as the current and anticipated business applications.

Definition of an Entity - A unique combination of name and location conducting business within the agricultural industry

Definition of Duplicate Entity - An entity with the same name and location combination as another entity within AGIIS

These definitions are at the core of our duplication prevention and identification logic.

Duplicate Prevention Approach

The approach begins with identifying the sources of incoming data from subscribers and then using the appropriate tools to identify and prevent duplicates. AGIIS currently receives entity Adds and Updates from the website, web services and bulk processes. AGIIS also has an annual de-duplication process that has been developed to ensure the rules are being enforced throughout the system equally and catch any duplicates that may have been introduced throughout the year in maintenance activities. The tool set used to prevent duplicates is consistant regardless the source of the incoming data. A high level diagram that identifies the different sources of data, and how it goes through a duplicate detection process before it enters AGIIS, is contained in Appendix A.

At the center of our solution is address standardization/matching software. AGIIS utilizes First Logic, a third party Coding Accuracy Support System (CASS) certified software product. The software initially standardizes the input (name and address) provided and then uses a matching algorithm to determine an appropriate match. This ensures that the input fields we are matching on can find a match on AGIIS. For example, if 123 Adams Street was submitted, the process would standardize the address to be 123 Adams ST, which is what would be displayed in the database if there was a match. Each data element that is used as part of the matching process has been configured within the software and a match level has been assigned to determine an acceptable tight or fuzzy match. These match levels were set by the Directory Oversight Committee after reviewing sample data sets and making sure they closely match the entity definition established. When entities are processed through the system, they are assigned a composite numeric score which determines if the entity matches an existing entity within AGIIS. This software and functionality is used throughout the application for standardization and duplicate detection. We also leveraged this software in the annual de-duplication process (See Appendix B Duplicate Matching Logic workflow).

Website and Web Services

The processes that are in place on the website and web services (i.e. places where real time performance is required) provide a comprehensive approach to detect and prevent duplicates from entering the database using a combination of automated tools and human intervention. The AGIIS website is the primary tool used to perform one-off entity Adds and Update requests. When a user submits an Add or Update request with a similar name and address information to an existing record in AGIIS, the “Potential Duplicate List” page will appear (Appendix C). The Potential Duplicate List page displays possible duplicate entity records that match or are similar to the entity attempting to be submitted. The duplicate criterion consists of both soundex and “like” queries on the name and a full address match. If an existing record has an exact match on the address and a name that "sounds like" the name on the request, the existing record will be highlighted in red for easy identification.

A warning message will display: *"Entities exist that exactly match the address and sound like the name of the entity you are about to add. Please carefully review the highlighted entities in the list to*

make sure your new entity does not already exist in AGIIS. Note that clicking the Continue button will result in the review of your Add request by the Directory Administrator (Please note that the Add request will no longer be automatically added!)" (Appendix D). If the user determines that the highlighted entity is the entity they want, they click on the entity name to bring up the View Entity page where the record can be added to the user's subset. If the user determines that the highlighted entity is NOT the entity they want; they would click the Continue button. Instead of the record being added immediately to the directory, an entity request will be sent to the Member Services for review. Before the request is reviewed by Member Services it is run through an automated process, using the matching algorithm from the bulk process, which provides a more extensive review of the request. If the automated process determines that the requested entity is, after all, unique, the record will be added to AGIIS. If the automated process determines that the request is a match to an existing entity, the request will be manually reviewed by the AGIIS Help Desk. The Help Desk will visually inspect record and perform internet searches if needed. If the request is unique, the record will be added to AGIIS. If the Help Desk determines that the requested entity is a duplicate of an existing entity record, the request will be declined. An email will be sent to the user with the results of the request.

Bulk Processes

Entity Bulk processes represents the transactional processing of entity records in batch mode. Subscribers can submit transaction files to perform the following entity functions: Add, Update, Report Duplicates, Re-activations, Add-to-Subset, Delete-from-Subset, Lookup by Demographics, and Lookup by Identifier. The transaction file is processed in "bulk" and an analogous return file is created which consists of the input data for each transaction with result data (e.g. return code, error code, GLN Identifier) appended to the end of each record. During the batch process the address standardization/matching software described in the Duplicate Prevention approach is leverage to prevent duplicates from entering the system (Appendix E shows the work flow of the batch process and where address standardization/matching software, First Logic, is used for duplicate prevention). There is also an extensive list of Return and Error codes available from the links menu on the AGIIS website that provides subscribers a list of why a record may have been added or rejected.

Duplicates Certification and Reporting

The Duplicates Certification and Reporting enhancement was implemented to allow AGIIS subscribers the ability to report suspected duplicate entities to Member Services for review. This process begins with a duplicates certification tutorial and quiz which is taken by a subscriber. Once successfully completing the tutorial and quiz, the user's AGIIS profile is updated giving the capability to report suspected duplicates. The Duplicates Certification process serves a dual purpose; it educates subscribers about AGIIS entity definitions and business rules and ultimately helps reduce the frequency of duplicate entities in the directory. If you are interested in learning more about the difference between a unique entity and a duplicate entity click on Duplicate Certification.

Once the user is certified, they can report suspected duplicates using the web interface, bulk process

or web services. An example of how duplicates are reported through the website is provided in the Duplicate Certification Training. Member Services reviews each submission to determine whether or not the entities meet the AGIIS definition of a duplicate. If so, the system will determine which entity will become the surviving entity, and which one will be considered the duplicate, based on the approved set of survivorship rules (Appendix F). The system also has data harvesting functionality built in, which allows the administrator to harvest information from subordinate records and populates the data into the surviving record if it is not already present (Appendix G).

Annual De-duplication Process

This is an activity that usually occurs towards the end of the calendar year, on a weekend, so normal business processes are not disturbed. The de-duplication process is performed after other annual maintenance activities such as Locatable Address Conversion System Updates (LACS) and the National Change Of Address Updates (NCOA) so if any duplicates are generated by an update they are resolved by the process.

To perform the de-duplication process, the established matching routines that are configured in the matching software are leveraged. Only active entities are eligible for the deduplication process. As duplicates are identified throughout process, a survivorship prioritization process has been identified so the “best” record can be deemed as the surviving record.

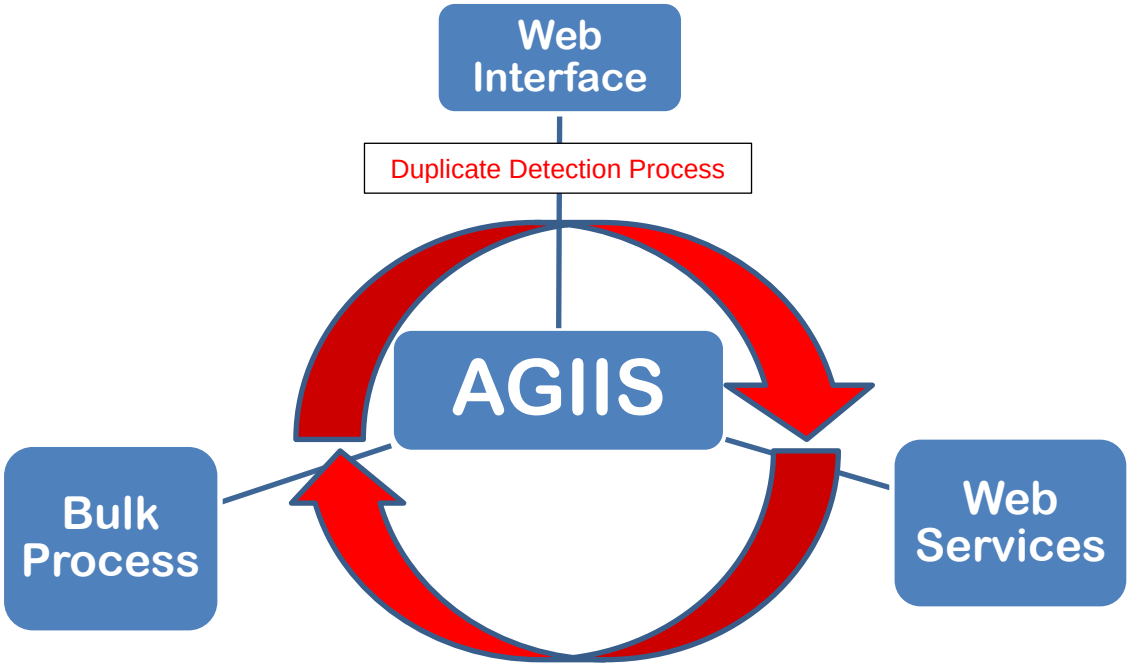
Survivorship prioritization is the process of determining which record should survive when duplicates are present. The same survivorship priority utilized during the annual de-duplication process is also used by Member Services when processing one-off duplicate requests (See Appendix F).

The de-duplication process also includes a Data Harvesting component. Data harvesting is the process of copying data from a record being replaced into the survivor record. This occurs when there are multiple records in a duplicate set and the survivor record lacks data that exists on a subordinate record (e.g. phone number). Extreme caution needs to be taken when performing data harvesting so that the meaning of the data is not falsely altered. The same Data Harvesting functionality is used by Member Services when processing one-off duplicate requests. The fields that are candidates for data harvesting can be found in Appendix G.

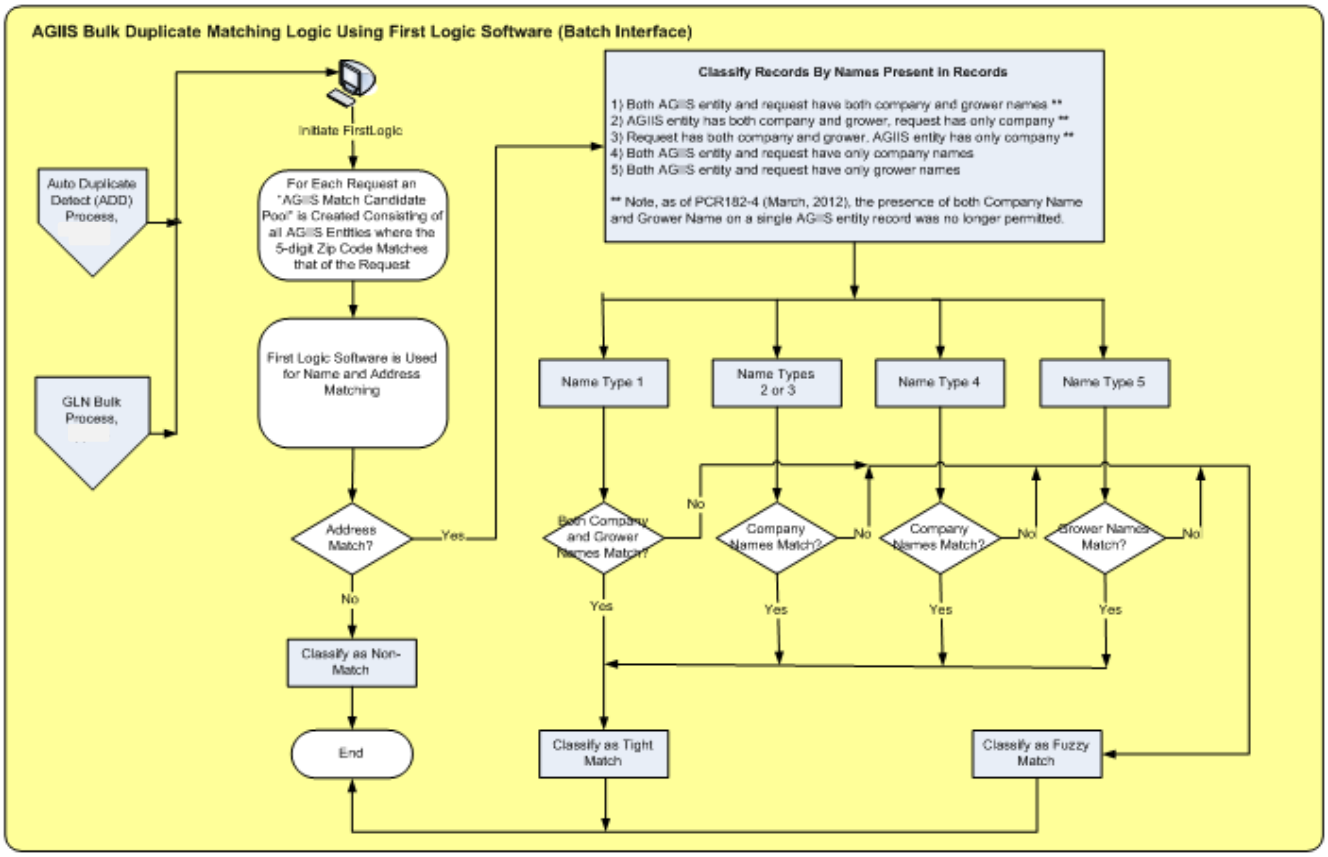
Once the process is complete, subscribers will receive a file containing entities in their subset that were inactivated due to the de-duplication process along with the identifier of the surviving record. It is recommended that each subscriber review and apply these updates so they can stay synchronized with AGIIS.

If you have questions regarding duplicate prevention or about how to become certified to report duplicates please contact Member Services at (866) 251-8618 or at Member.Services@AgGateway.org.

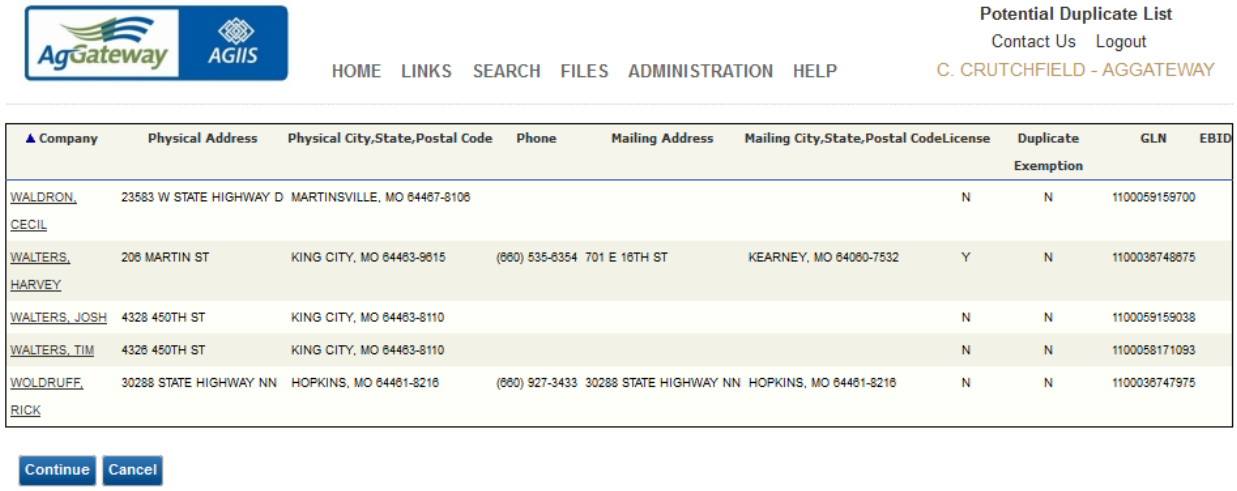
Appendix A


Web Interface
Duplicate Detection Process
AGIIS
Bulk Process
Web Services

Appendix B


AGIIS Bulk Duplicate Matching Logic Using First Logic Software (Batch Interface)
Initiate FirstLogic
Auto Duplicate Detect (ADD) Process,
GLN Bulk Process,
For Each Request an "AGIIS Match Candidate Pool" is Created Consisting of all AGIIS Entities where the 5-digit Zip Code Matches that of the Request
First Logic Software is Used for Name and Address Matching
Address Match?
Yes
No
Classify as Non-Match
End
Classify Records By Names Present in Records
1) Both AGIIS entity and request have both company and grower names **
2) AGIIS entity has both company and grower, request has only company **
3) Request has both company and grower. AGIIS entity has only company **
4) Both AGIIS entity and request have only company names
5) Both AGIIS entity and request have only grower names
** Note, as of PCR182-4 (March, 2012), the presence of both Company Name and Grower Name on a single AGIIS entity record was no longer permitted.
Name Type 1
Name Types 2 or 3
Name Type 4
Name Type 5
Both Company and Grower Names Match?
Company Names Match?
Company Names Match?
Grower Names Match?
No
Yes
Classify as Tight Match
Classify as Fuzzy Match

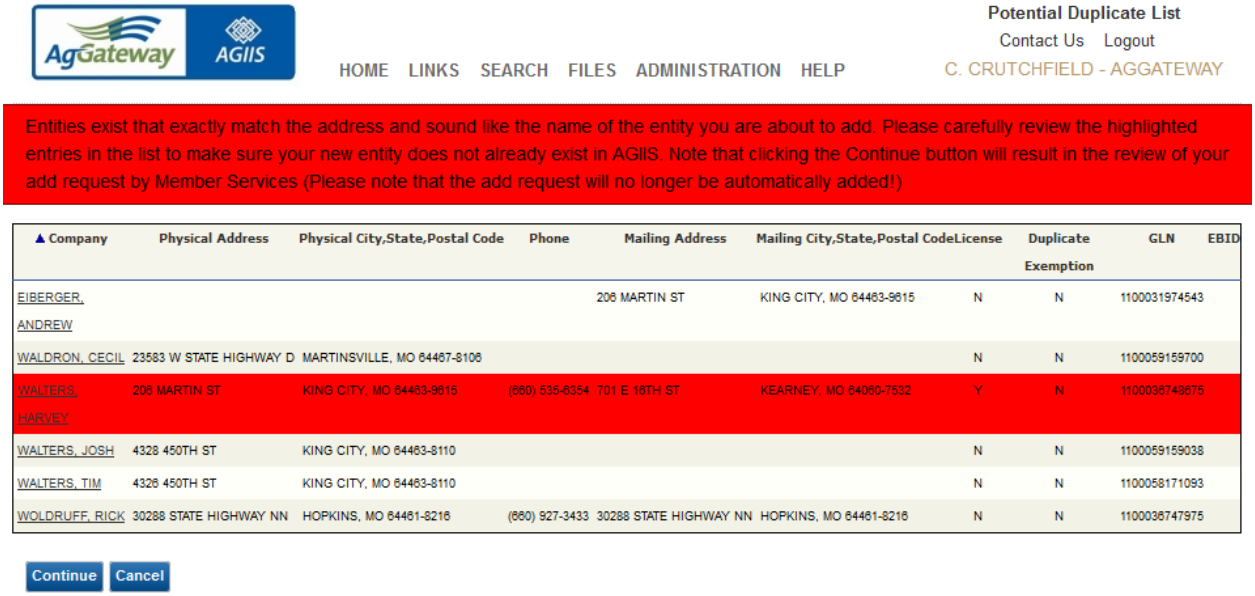
Appendix C



▲ Company	Physical Address	Physical City,State,Postal Code	Phone	Mailing Address	Mailing City,State,Postal Code	License	Duplicate Exemption	GLN	EBID
WALDRON, CECIL	23583 W STATE HIGHWAY D	MARTINSVILLE, MO 64467-8106				N	N	1100059159700	
WALTERS, HARVEY	206 MARTIN ST	KING CITY, MO 64463-9615	(660) 535-6354	701 E 16TH ST	KEARNEY, MO 64060-7532	Y	N	1100036748675	
WALTERS, JOSH	4328 450TH ST	KING CITY, MO 64463-8110				N	N	1100059159038	
WALTERS, TIM	4326 450TH ST	KING CITY, MO 64463-8110				N	N	1100058171093	
WOLDRUFF, RICK	30288 STATE HIGHWAY NN	HOPKINS, MO 64461-8216	(660) 927-3433	30288 STATE HIGHWAY NN	HOPKINS, MO 64461-8216	N	N	1100036747975	

Continue Cancel

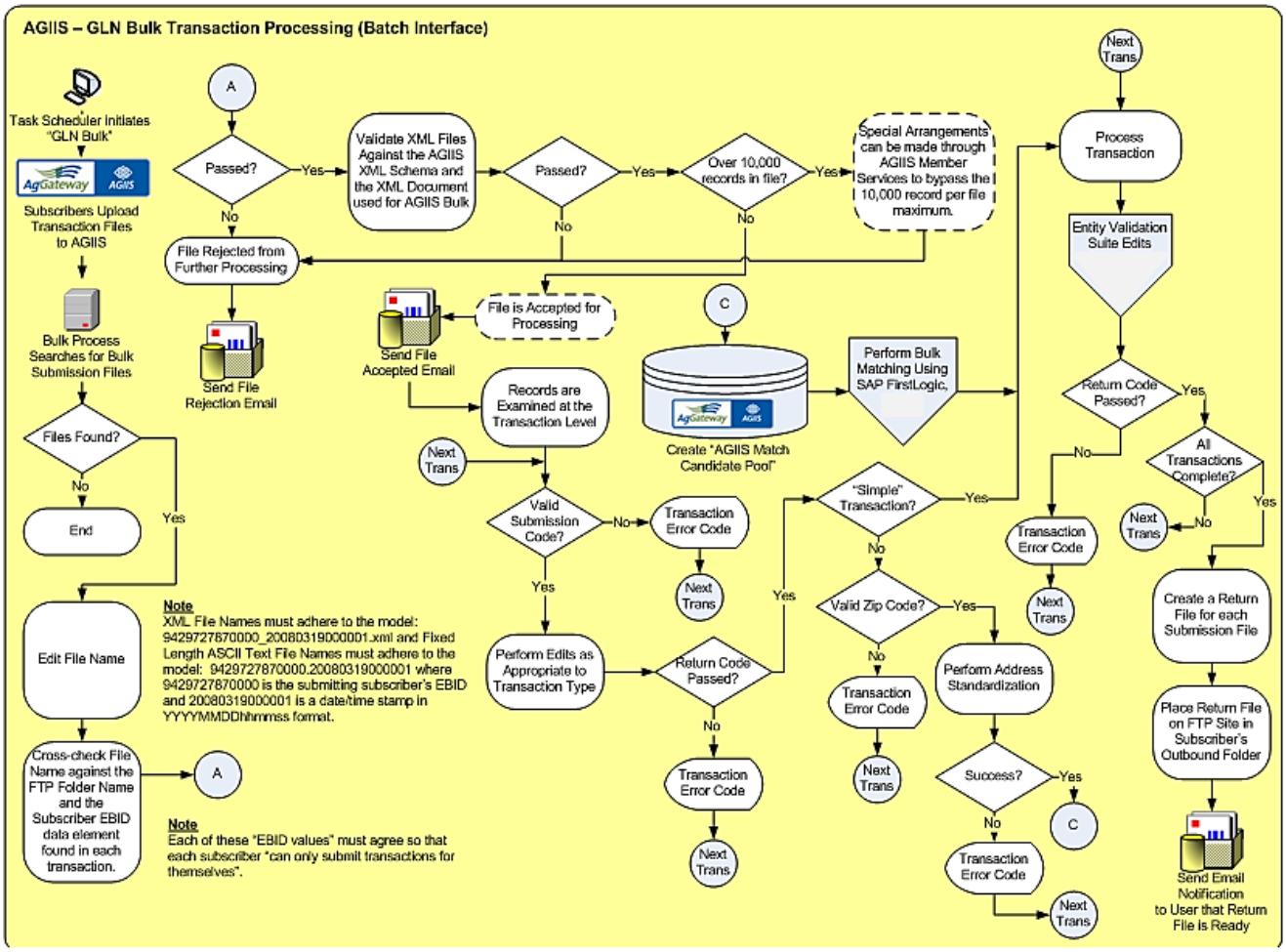
Appendix D



▲ Company	Physical Address	Physical City,State,Postal Code	Phone	Mailing Address	Mailing City,State,Postal Code	License	Duplicate Exemption	GLN	EBID
EIBERGER, ANDREW				206 MARTIN ST	KING CITY, MO 64463-9615	N	N	1100031974543	
WALDRON, CECIL	23583 W STATE HIGHWAY D	MARTINSVILLE, MO 64467-8106				N	N	1100059159700	
WALTERS, HARVEY	206 MARTIN ST	KING CITY, MO 64463-9615	(660) 535-6354	701 E 16TH ST	KEARNEY, MO 64060-7532	Y	N	1100036748675	
WALTERS, JOSH	4328 450TH ST	KING CITY, MO 64463-8110				N	N	1100059159038	
WALTERS, TIM	4326 450TH ST	KING CITY, MO 64463-8110				N	N	1100058171093	
WOLDRUFF, RICK	30288 STATE HIGHWAY NN	HOPKINS, MO 64461-8216	(660) 927-3433	30288 STATE HIGHWAY NN	HOPKINS, MO 64461-8216	N	N	1100036747975	

Continue Cancel

Appendix E


AGIIS – GLN Bulk Transaction Processing (Batch Interface)
Task Scheduler Initiates "GLN Bulk"
Subscribers Upload Transaction Files to AGIIS
Bulk Process Searches for Bulk Submission Files
Files Found?
No
End
Yes
Edit File Name
Cross-check File Name against the FTP Folder Name and the Subscriber EBID data element found in each transaction.
A
Note
Each of these "EBID values" must agree so that each subscriber "can only submit transactions for themselves".
A
Passed?
No
Yes
File Rejected from Further Processing
Send File Rejection Email
Validate XML Files Against the AGIIS XML Schema and the XML Document used for AGIIS Bulk
Passed?
No
Yes
Over 10,000 records in file?
No
Yes
Special Arrangements can be made through AGIIS Member Services to bypass the 10,000 record per file maximum.
File is Accepted for Processing
Send File Accepted Email
Records are Examined at the Transaction Level
Next Trans
Valid Submission Code?
No
Yes
Transaction Error Code
Next Trans
Note
XML File Names must adhere to the model: 9429727870000_20080319000001.xml and Fixed Length ASCII Text File Names must adhere to the model: 9429727870000.20080319000001 where 9429727870000 is the submitting subscriber's EBID and 20080319000001 is a date/time stamp in YYYYMMDDhhmmss format.
Perform Edits as Appropriate to Transaction Type
Return Code Passed?
No
Yes
Transaction Error Code
Next Trans
C
Create "AGIIS Match Candidate Pool"
Perform Bulk Matching Using SAP FirstLogic.
"Simple" Transaction?
No
Yes
Valid Zip Code?
No
Yes
Transaction Error Code
Next Trans
Perform Address Standardization
Success?
No
Yes
C
Transaction Error Code
Next Trans
Next Trans
Process Transaction
Entity Validation Suite Edits
Return Code Passed?
No
Yes
Transaction Error Code
Next Trans
All Transactions Complete?
No
Yes
Next Trans
Create a Return File for each Submission File
Place Return File on FTP Site in Subscriber's Outbound Folder
Send Email Notification to User that Return File is Ready

Appendix F

Survivorship Rules
Survivorship prioritization is the process of determining which record should survive when duplicates are present. The same survivorship priority will be utilized during the annual de-duplication process.

1. Active AGIIS Subscriber record
2. Subscriber Owned GLN
3. Non Subscriber Owned GLN
4. Last Verification - Self Verified
5. Last Verification - Member Services Verified Tier 2
6. Last Verification - Member Services Verified Tier 1
7. Last Verification – Subscriber Verified
8. Last Verification Date
9. Licensed Yes
10. Subset Count
11. Entity Creation Date (Oldest survives.)

Appendix G

Data Harvesting
Data harvesting is the process of copying data from a record being replaced into the master/survivor record. This occurs when there are multiple records in a duplicate set and the survivor record lacks data that exists on a subordinate record (e.g. phone number). The following fields are candidates for data harvesting in the automated de-duplication process:

- ❑ Phone number
- ❑ Latitude
- ❑ Longitude
- ❑ Standard Point Location Code (SPLC)