AHGF (geofabric) production – an analysis of processes, descriptions and models

Box, P and Atkinson, R

June 2011

Report to Bureau of Meteorology

A water information R & D alliance between the Bureau of Meteorology and CSIRO’s Water for a Healthy Country Flagship
Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills.

CSIRO initiated the National Research Flagships to address Australia’s major research challenges and opportunities. They apply large scale, long term, multidisciplinary science and aim for widespread adoption of solutions. The Flagship Collaboration Fund supports the best and brightest researchers to address these complex challenges through partnerships between CSIRO, universities, research agencies and industry.

The Water for a Healthy Country Flagship aims to provide Australia with solutions for water resource management, creating economic gains of $3 billion per annum by 2030, while protecting or restoring our major water ecosystems.

The work contained in this report is collaboration between CSIRO’s Water for a Healthy Country Flagship and the Bureau of Meteorology’s Water Division under the Water Information Research and Development Alliance (WIRADA).

WIRADA is a five-year, $50 million partnership that will use CSIRO’s R&D expertise to help the Bureau of Meteorology report on availability, condition and use of Australia’s water resources.

For more information about the partnership visit www.csiro.au/partnerships/WIRADA.html
For more information about Water for a Healthy Country Flagship or the National Research Flagship Initiative visit www.csiro.au/org/HealthyCountry.html

Citation: Box, P and Atkinson R. 2011. AHGF (Geofabric) production – an analysis of processes, descriptions and models. CSIRO: Water for a Healthy Country National Research Flagship

Copyright and Disclaimer
© 2011 Bureau of Meteorology and CSIRO. To the extent permitted by law, all rights are reserved and no part of this publication covered by copyright may be reproduced or copied in any form or by any means except with the written permission of Bureau of Meteorology and CSIRO.

Important Disclaimer
Bureau of Meteorology and CSIRO advise that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, Bureau of Meteorology and CSIRO, including each of its employees and consultants, excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Cover Image
Created by Paul Box. Copyright © 2012 CSIRO
ACKNOWLEDGMENTS

This work is part of the water information research and development alliance between CSIRO’s Water for a Healthy Country Flagship and the Bureau of Meteorology.
EXECUTIVE SUMMARY

This report describes a gap analysis looking at the differences between actual production process to create AHGF data products, the descriptions of the process in the data product specification, and the mapping between models of the product and the sources they were derived from. The analysis focuses on the production of SH_network product from the maintenance database and is largely driven by the longer term need and research agenda related to executable data product specifications (xDPS). The report also highlights the current status of model driven approaches and their role in Geofabric design and production. This is intended to inform discussion about the development of data product specifications, and application of Solid Ground tools in the SWIM 2011/12 project, in relation to geofabric production, and as part of a broader discussion of the role of MDA and Solid Ground in AWRIS.

From a model perspective, Phase 1 data product schemata were generated manually and validated against the conceptual model. As part of Phase 2 of the Geofabric processes to augment Phase 1 products and to identify stable hydrologic concepts needed to realise the business objectives of the Geofabric are being developed. In Phase 1, the schema generation process was based on, but not driven by models. UML models for this analysis were reverse engineered from geodatabase schemas for the maintenance and product models. Using a range of Solid Ground functionality, models were imported into Enterprise Architect, cleaned (to strip out system metadata e.g. common geofabric attributes), converted to ISO compliant models, and re-factored to remove the ESRI meta-model elements and extract common vocabularies. This (reverse engineering) process of generating these models and the mappings between them is the reverse sequence of steps that would be performed if end to end model driven geospatial product development were implemented. However, this reverse engineering process greatly informs the requirements for forward engineering transformations.

The actual process to create the Phase 1 version of the SH_network data products was largely manual and no scripts or custom processes were used. The data product is in effect a subset of features in the maintenance database with minor modifications to join feature attributes with those in a table.

Within data product specifications, lineage metadata elements (from ISO 19115) are used to describe the processes and input data used to create products. This mechanism is severely limited as it is focused on metadata as data sets. The production processes for SH_network, as is increasingly typical for environmental geospatial products, are complex, spanning organisations, computing environments and applications with highly variable quality and level detail of metadata. Thus, it is not feasible to capture this lineage metadata (or provenance information) manually and almost impossible to adequately describe in textual form in any kind of meaningful way.

Currently, models are not used to drive production processes and thus models represent informative metadata. Consequently, there is a disconnect between any model mapping that can potentially be used to drive production and the actual production processes. It is the underlying objective of the model driven approach to support and drive aspects of geospatial information production using models. However, to achieve the potential benefits of specification-driven data production it is necessary to:

- understand the Bureau’s requirements for and scope of models and mapping use to drive operational geospatial data production;
- develop methodologies and how-to guides that enable consistent, manageable approaches to modelling and model mapping. In particular:
  - approaches to refactoring and extracting/abstracting out platform specific and other implementation specific details
  - approaches to use of mappings; and
- approaches to referencing geospatial processes in model mappings that transform feature types. For example a line generalisation applied as part of the transformation in a mapping from a cartographic stream segment feature class in the maintenance model to a generalised representation of the line in a derived product.
In addition to the use of models to drive geospatial production, the resultant models and mappings between them represent critical system metadata, which can be used at run-time (‘actioned’) as ‘operational metadata’. The clear articulation of use cases for exploiting models and their mappings, will greatly inform the scope of and approaches to mapping models and model elements to each other. SWIM project 2011/12 activity 2 will focus on articulating a registry exploitation use case related to a feature type catalogue.

In SWIM project 2011/12, the Hydrologist’s WorkBench (HWB) system will be used to provide a mechanism for the generation and evolution of Geofabric products. The processing algorithms and business rules developed by Bureau Geofabric team will be ported to HWB for execution. The use of HWB will enable continued experimentation to refine existing data products, development of new products and research leading to automation of product updates. In addition, the use of HWB and the mining of workflow metadata to create provenance information about data products, builds towards the long-term research agenda of executable data product specifications.
1. INTRODUCTION

1.1. Overview

This report describes a gap analysis looking at the differences between actual production process to create AHGF data products, the descriptions of the process in the data product specification for data products, and the mapping between models of the product and the sources they were derived from. The primary objective of this gap analysis to identify business rules, processes and logic used to generate these products that have not been captured as formal models (conceptual, implementation or mappings between these) or described in data product specifications. Please refer to the geofabric methodology paper for a detailed description of the role of models and data product specifications in the geofabric (Atkinson, 2009).

This analysis will be used to identify aspects of the production process not currently described or modeled which need to addressed in order to automate production processes. In addition the analysis is being used to identify critical gaps in and inform further development of model mapping capabilities in the context of ongoing maintenance and evolution of the AHGF design and its integration with other systems.

This analysis is largely driven by the research agenda related to executable data product specifications (xDPS). This long term research agenda is investigating automation of end user access to aspects of the Geofabric, through the articulation of inputs and processes using a parameterised data product specification. In order to achieve this following is required:

- A maintenance environment to store geofabric features that will be used to produce products (the data inputs) together with models of features types in the maintenance environment;
- A means to be able to adequately describe and reference processes required to produce products from the maintenance environment (the processes).

The ability to specify data inputs and processes performed on them is a significant pre-requisite for automating data production. It also has value as documentation and provenance record for both end users and long term maintenance of the Geofabric. Therefore, the analysis looks at what these data inputs and processes are in reality and how they are described in the data product specification.

1.2. Geofabric production processes at phase 1

The Australian Hydrological Geospatial Fabric (AHGF) is ‘a consistent representation of water features, and their connectivity, in the Australian water system’ (Smith, Dee et al., 2009) It is underpinned by an implementation independent conceptual model representing those features that can be consistently referenced within the Australian hydrological context and the relationships between these features – the AHGF Conceptual Model. The purpose of this model is to assist in managing the hydrological concepts and the relationships between these concepts. In addition, it is intended to document the relationships between concepts and derived products in which the concepts are delivered as features. This may be used to cross reference alternative representations of these features found in different contexts. This concept is reflected in the Identifier Best Practice review (Cox, 2011), which highlighted the emerging best practice to separate concepts and representations with distinct identifiers. The model links these together.

1.2.1. AHGF maintenance database

Version 1 of the AHGF conceptual model was delivered by the SWIM project in Year 2 of the project (09-10). A schema for the maintenance database to persist geofabric features was developed manually and checked for consistency and conformance against the AHGF conceptual model.

The maintenance database, implemented as an ESRI geodatabase, stores AHGF features and the relationships them, in conformance with the conceptual model. The maintenance database is not exposed to end users, but is used as a staging area for features which are composed into data products. The maintenance database however is an instantiation of parts of the AHGF conceptual model necessary to support then generation of derived phase 1 products, rather than being a complete representation of it.
Figure 1 below, shows the simplified structure of the AHGF maintenance geospatial database. Each feature class is shown along with an indication of the populating data sources in order to indicate how AusHydro and ANU deliverables relate to the schema. Raster datasets, including both source and derived products are also represented.

1.2.2. Populating the maintenance database

The maintenance database was populated based on supplied data products - AusHydro (provided by GA) and network streams and catchments provided by (ANU). These supplied products were processed, and ‘recomposited’ and loaded into the maintenance database. In phase 1 the geoprocessing workflow to process input features, create new derived features and to be loaded into the database geofabric is long and complex and revolves around the identification of contracted nodes and associated contracted catchments. In phase 2 the process has been significantly enhanced to address a number of significant deficiencies in the scope of contracted nodes and the resultant contracted catchment coverage produced.

1.2.3. Phase 1 products

The AHGF maintenance database was used as the basis for the generation of the following phase 1 data products:

- SH_cartographic
- SH_catchments
- SH_network

These products were designed following consultation with end users. The schemas for the products were developed manually (using Visio), and were validated for conformance against the maintenance database model, which is an instantiation of relevant aspects of the AHGF conceptual model. The products were released in Oct 2010 and delivered as ESRI geodatabases based on the...
schema. These phase 1 products represent initial ingestion, recompilation and the production of initial products. Limitations of these products are discussed in 3.1.1.

Data Production Specifications were produced for each product in accordance with a template based on ISO 19131 Data Product Specifications. A template and populated examples for geofabric products were deliverables in SWIM 09-10\(^1\). Work is currently nearing completion to implement the processes that will enable the creation of phase 2 derived geofabric products.

1.3. The role of models in product design and production

Geofabric data exists in both the maintenance environment and the individual products in which data is delivered. The products implement the same feature in different forms, and exhibit a lot of necessary duplication of features and attributes. A conceptual view of these implementation models is used to document and enforce consistency between geofabric products delivering different permutations of geofabric features. One focus of the SWIM project has been on tools to allow the platform specific product models and a more conceptual view to be kept in alignment, through a process known as Model Driven Engineering (MDE). MDE supports reverse-engineering and schema generation. This has been extended with tools to enable modular information models to be created at the conceptual level and mapped to each implementation to define the relationships that exist and to support system interoperability.

Mappings between models play a significant role in mapping between layers of abstraction and to support transformations between different yet consistent views of features across products. Firstly, mappings enable the tracing of an abstract concept such as ‘Basin’, (in the AHGF conceptual model) to concrete feature types (and associated description) such as ‘AHGF_ContractedCatchment implemented in the geofabric persistence environment.

Secondly, the mappings may describe transformations between features in the maintenance environment and feature delivered in a product. This may be a simple transformation or calculation such as changing an attribute name, or performing a calculation on attribute values. It may also be a more complex transformation or calculation e.g. a line generalisation geo-process or the creation of a new derived feature type from one or more input feature types. Having these mappings modelled, it will be possible to, for example, find alternative views or observation data related to a specific feature. An example of the potential is:

- A feature (e.g. “Lake Pieman” from the GeoFabric SH_Cartographic product) is referenced in some context, such as a map on a web portal;
- This feature is an instance of a particular feature type (AHGF_SH_Carto:Reservoir)
- AHGF_SH_Carto:Reservoir is a representation of a concept HY_Features:Storage (a more abstract feature type)
- HY_Features:Storage is also represented in another system by SLAKE:Storage
- SLAKE:Storage is realised by the slake_storages controlled vocabulary in the slake system (which uses a different form of identifier
- Slake_storages is used in the dataset SLAKE_water_level_observations
- The user is able to find and traverse these relationships, and hence ask for water level observations for Lake Pieman
  - The “Lake Pieman” is used to find the common identifier F
- F is used to find the alternative identifier urn:used in slake_storages
- SLAKE_water_level_observations is queried using the specific identifier it understands to find the related observations

\(^1\) Template and populated example for SH_Cartographic product are available on WIRADA sharepoint.
Whilst this may seem somewhat convoluted, it is easily automatable and addresses the very complex problem of determining how systems interoperate when they refer to the same object using different identifiers because they use different representations of the same thing.

To enable this resolution the following is required:

- relationships between features be defined in the layers of information models
- relationships are exposed at the level of both feature types and feature instances through service interfaces

In this context, feature type information may be regarded as a type of “Feature Type Catalog” (FTC) and it appears that there is strong interest in a standardised, model driven FTC using a Linked Open Data (identifier oriented) interface.
2. REVIEW OF PRODUCTION PROCESS, MODELS AND DESCRIPTIONS

2.1. Scope and methodology

The analysis described in this document focuses on the SH_network product and its production from the AHGF maintenance database. In this context the role of mappings to define and support execution of transformation between different views of features types is the focus so the model mapping aspect of this analysis. The analysis comprises the following steps:

- Review and documentation of Phase 1 Surface Hydrology Network Product (SH_network) production process;
- Review of the data product specifications and their description of the processes;
- Review model and model mappings from maintenance to product models; and
- Analysis of inconsistencies between the actual process and processes recorded in DPS and represented in models.

2.2. Production process

Phase 1 Geofabric products were developed in the following manner. Firstly, the AHGF conceptual model was used to develop an understanding of how the data products would relate to each other, essential providing design constraints. Implementation oriented UML diagrams for the products were developed using ESRI case tools provided in Microsoft Visio. This produces a visual model, but does not create a fully articulated and extensible information model beyond the diagrams. (In modelling environments the diagrams are views of the model, and multiple views may be kept synchronised with a single model). A single diagram was developed for all data products. Models were then exported to XML Metadata Interchange (XMI) format and using ArcCatalog, geodatabase schemas were developed for product geodatabases from the XMI files.

Several manual processes were then applied to the resultant geodatabases, to create relationships between feature class attributes and a table containing contracted nodes. Finally, the ArcCatalog load wizard, was used to map fields from maintenance database to the product databases, and to load product geodatabases with content from the maintenance database.

The Microsoft Visio diagramming environment was used, in spite of the limitations that required manual completion of the schema design, as the functionality to support the development of ESRI UML models was not at the time sufficiently mature in alternative environments. ESRI itself has ceased to develop the Visio platform as a CASE tool due to its inherent limitations. The experience was used to design a function with an improved capability in the Solid Ground toolset. This capability has since been adopted and developed as a built-in function in the Sparx systems Enterprise Architecture product.

2.3. SH_network Data product specification

Data product specifications for geofabric products were produced based on a template (Box, O’Hagan et al., 2009) which in-turn was based on ISO standard 19131 Data Product Specifications (ISO, 2008). The data product specification (DPS) developed for SH_network (Geofabric Management Unit, 2010) provides, in section 6.2.1 Lineage statement, a high-level description of the data sources and production process and the SH network product. The following processing steps are defined in the data product specification:

- Step 1: ANUDEM Streams dataset is received and loaded into the Geofabric development GIS environment.
- Step 2: Feature classes from ANUDEM Streams are recomposed into composited Geofabric Feature Dataset Feature Classes in the Geofabric Maintenance Geodatabase.
- Step 3: Re-composited feature classes in the Geofabric Maintenance Geodatabase Feature Dataset are assigned unique Hydro-IDs using ESRI ArcHydro for Surface Water (ArcHydro: 1.4.0.180 and ApFramework: 3.1.0.84).

Step 4: Feature classes from the Geofabric Maintenance Geodatabase Feature Dataset are extracted and reassigned to the Geofabric Surface Network Feature Dataset within the Surface Network Geodatabase. This represents a description of the primary source of input feature types to the Geofabric i.e. ANUDEM streams, the re-composition of features into the AHGF maintenance database, and are re-composition of these features into the SH_network product. As part of the derived product creation process, a number of new feature classes were created. Table 1, from the Geofabric Product Guide (Bureau of Meteorology, 2010), which is referenced in the SH_network DPS, lists features comprising the network product. New features created by the Bureau are highlighted in blue.

<table>
<thead>
<tr>
<th>AGHF_Network – Feature Class.Subtype(Type)</th>
<th>Feature Class Geometry</th>
<th>AHGF Feature Type Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHGFNetworkSegment.FlowSegment</td>
<td>line</td>
<td>1</td>
</tr>
<tr>
<td>AHGFNetworkSegment.ArtificialFlowSegment</td>
<td>line</td>
<td>2</td>
</tr>
<tr>
<td>AHGFNetworkSegment.WaterAreaSegment</td>
<td>line</td>
<td>3</td>
</tr>
<tr>
<td>AHGFNetworkNode.NetworkJunctionNode</td>
<td>points</td>
<td>4</td>
</tr>
<tr>
<td>AHGFNetworkNode.NetworkTerminusNode</td>
<td>points</td>
<td>5</td>
</tr>
<tr>
<td>AHGFNetworkNode.NetworkArtificialNode</td>
<td>points</td>
<td>6</td>
</tr>
<tr>
<td>AHGFNetworkNode.NetworkWaterAreaNode</td>
<td>points</td>
<td>7</td>
</tr>
<tr>
<td>AHGFNetworkNode.NetworkGhostNode</td>
<td>points</td>
<td>8</td>
</tr>
<tr>
<td>AHGFNetworkNode.NetworkHeadNode</td>
<td>points</td>
<td>9</td>
</tr>
<tr>
<td>AHGFNetworkNode.NetworkCliffNode</td>
<td>points</td>
<td>10</td>
</tr>
<tr>
<td>AHGFCArcher.Catchment</td>
<td>polygon</td>
<td>21</td>
</tr>
<tr>
<td>AHGFCArcher.Catchment.ContractedArea</td>
<td>polygon</td>
<td>22</td>
</tr>
<tr>
<td>AHGFCArcher.Catchment.NonContractedArea</td>
<td>polygon</td>
<td>23</td>
</tr>
<tr>
<td>AHGFCArcher.Catchment.NoFlowArea</td>
<td>polygon</td>
<td>24</td>
</tr>
<tr>
<td>AHGFWaterbody.Reservoir</td>
<td>polygon</td>
<td>25</td>
</tr>
<tr>
<td>AHGFWaterbody.Lake</td>
<td>polygon</td>
<td>26</td>
</tr>
<tr>
<td>AHGFWaterbody.Swamp</td>
<td>polygon</td>
<td>27</td>
</tr>
<tr>
<td>AHGFNetworkSegment_FS</td>
<td>na</td>
<td>table</td>
</tr>
</tbody>
</table>

Table 1 Feature type registry for Geofabric Surface Network product

### 2.4. Model creation and mapping process

As described in section 1.2, models used to generate the schemas for the maintenance and the product databases were created in Visio case tools. Rather than attempting to convert these models to UML models in EA and deal with the idiosyncratic handling of UML in Visio for the ESRI case tool, it was decided that the UML models for this analysis would be reverse engineered from the geodatabase schemas. This allows the faithful capture of relationships and implementation metadata in the UML model, providing an improved capability over the simple functions supported in the Visio environment.

An XML workspace document (an export file containing, inter-alia, the schema for a given geodatabase) was generated for each product geodatabase. The XML schema for each geodatabase was then imported into a UML workspace (using Solid Ground’s “import ArcGIS Workspace” function).

These models were then refactored to:

- Identify then strip out common “housekeeping” attributes used in all feature types and thus considered to be part of a geofabric implementation “system metadata” model). This system

---

metadata model is itself captured in UML and can be automatically re-applied during schema generation time (this is future functionality). (Note that at this stage the governance of this system metadata model is not clear, however it can be registered in the model registry like any other model component. As the usage and governance becomes better understood, an appropriate set of relationships can be defined to relate this model to other models, and populated automatically during the refactoring process).

- Convert the model to a standard modelling language (ISO 19103) (“Convert UML Profile” tool)
- Remove the ESRI meta-model (Point, line, polygon superclasses) replacing these with spatial attributes. (This allows the actual semantics of the objects to be declared through inheritance, rather than only the typical geometry used for one aspect). This allows, for example, a lake to exist with both a boundary and a reference point. (This is currently a manual task but will be automated in future. There are subtleties regarding the semantics of geometric attributes in the ISO model that need further experience to understand how to choose these – for example the difference between a GM_SurfaceBoundary and a GM_SurfacePatch – both are probably represented as ESRI:Polygons, though MultiPolygons are used in some circumstances.)
- Abstract out common vocabularies and attributes. Elements common to more than 1 feature type within a product and those used across multiple products were extracted to common packages.

2.4.1. AHGF_Maintenance model refactoring
The maintenance model was refactored into the packages as per Figure 2, below. Key points of interest:

- «Application Schema» AHGF_Maintenance_AHGF_Framework contains the feature classes implicated in the geofabric networks and hierarchical catchments
- «Application Schema» AHGF_Maintenance_Hydrography - all other maintenance feature types
- «Application Schema» AHGF_Maintenance_Domains - contains all codelists for all feature types within the maintenance model
- «Application Schema» AHGF Common Vocabularies - contains vocabularies used within more than one product. This was produced by analysis of vocabularies used in the all three reverse engineered phase 1 product models.
Refactored maintenance model packages

Figure 3 below, shows the refactored AHGF_Maintenance_AHGF_Framework model. The framework package contains the key network feature types and thus the majority of types mapped to the SH_network product.)
2.4.2. **SH_Network product model refactoring**

The SH_network model was refactored into the following packages:

- Reuse of «Application Schema» AHGF Common Vocabularies - vocabularies used in more than one phase 1 product models
- «Application Schema» AHGF SH_Network Specific Domain Vocabularies – vocabularies used within the SH_network product
- «Application Schema» AHGF_SH_Network – the package containing the product model.

The refactored model for the AHGF_SH_Network is shown in Figure 4, below.
Figure 4 Refactored SH_network product model

2.4.3. Model mapping

Following refactoring of both models, feature types in the maintenance model were mapped to feature types in the product model. Since there is a strong correlation between Feature types (many of the same feature types are reused in different products), this can be efficiently done using a Solid Ground automated mapping function. This uses the feature type name in order to identify matches in source and target models. Note that any specific transformations can be manually added to this generated mapping to handle more complex cases if necessary. Automated mapping was performed on the entire SH_network model. Figure 5, shows the results of automated mapping for a single feature type - network streams.

AHGF (geofabric) production – an analysis of processes, descriptions and models V0.2
Key points to note with regard to the automated mapping:

- The mapping class maps attributes from the class in the maintenance model to attributes in the SH_network product class using the representsAttribute relationship between the mapping class and the maintenance class;
- Mapping to the target class (in the SH_network product model) is represented by a single mapsTo relationship with the mapping class;
- A number of attributes included in the maintenance model have been stripped out of the maintenance model as part of the refactoring process;
- This mapping, though trivial, provides a machine readable connection between the feature types in the two environments that is not declared in any other form; and
- It is not safe to depend on unique names at “run-time” to discover such relationships, but automated mappings with manual augmentation provides a simple and cost-effective means to make explicit relationships that can be actioned.
3. ANALYSIS

3.1. Overview

Before presenting an analysis of the differences between the actual production process and the representation of the production process in data product specification and models, it is worth articulating several relevant factors that impact on the analysis. These relate to:

- the nature of the products developed in phase 1 used as the basis for analysis
- the role of models and the nature of the modelling approaches used

3.1.1. Nature of phase 1 products

The products delivered in phase 1 do not implement key aspects of the conceptual model. The most critical product not yet produced, which represents the key concepts at the heart of the geofabric is the topological network product in a variety of different forms, built from contracted nodes and indexed using a stable hierarchical basin network (“contracted catchments”). Although a network product and associated catchments have been produced in phase 1, these products do not yet carry stable geofabric identity and do not include a range of supplemental contracted nodes developed in order to meet identified product requirements. These products are currently under development and will be delivered in phase 2. Key products are the full network built from contracted nodes and a simplified (dendritic) version of the network in a variety of representations including a topological (node link network) representation and a contracted catchment representation. These ‘core’ geofabric products carry and are used to maintain geofabric identity and relationships and are thus key internal maintenance products. However, they will also be delivered as end-user products. It is from these products that derived products will be developed. These include for example, the different aggregations of contracted catchments to create reporting a variety of regions to meet various reporting needs of the Bureau, and alternative simplified, filtered, subsets of the network for in different representation for different purposes.

Phase 1 products are currently simple extracts of the maintenance database, rather than being the products designed to meet the higher order business requirements of the Geofabric. Automated derivation of a suite of application-oriented products from the maintenance database was anticipated to be the focus of this research activity.

3.1.2. Maturity of model-driven capabilities

The approach used to create the models of the maintenance database and the SH network product and the mapping of these models, was the reverse sequence of steps that would be undertaken if end to end model driven geospatial product development were to be undertaken. The forward engineering sequence of steps that would be taken to design and implement product models would be, to use features from the maintenance model to create a product model and where necessary create new derived features by applying transformations described in model mappings, to existing features. This would result in the production of platform independent product (PIM) model. These models could then be used to generate ESRI (or other implementation) schemas by applying platform specific rules. This ‘forward engineering’ process is shown in Figure 6, below. A reverse engineering process was actually used to create models for this analysis. This process is also shown in Figure 6.

Research to-date has focused on mapping feature types across levels of abstraction to enable models and mappings between them to be traversed, thus enabling models to fulfill their destiny as operational system metadata. To-date, there has been insufficient context, business drivers or resources to focus on mapping to support transformation and instantiation of features as part of a geospatial production process. Therefore currently modelling is somewhat disconnected from the production process.
3.1.3. Key elements of the analysis

Given the limitations described above, related to the nature of products available for analysis, limitations in the maturity of the model driven process for geospatial product generation the analysis has focused on a simple case, the production of the SH_network product from the maintenance database.

From the perspective of model mapping, the analysis looks at the results of an automated mapping between feature at the same level of abstraction i.e. platform independent models of the maintenance and product models. Although this was a narrow focus, it did enable:

- Testing of aspects of the reverse engineering process
- Deeper understanding of requirements for stripping reverse engineered models of platform specific detail (extracting ESRI metamodels related elements)
- Understanding requirements for refactoring the reverse engineered models to pull out common components (such as shared attributed and vocabularies (implementation details))

3.2. Geofabric production processes

The actual process to create the SH_network data product was largely manual and no scripts or custom processes were used. The data product is in effect a subset of features in the maintenance database with minor modifications to join feature attributes with those in a table. Currently all
products are delivered as ESRI geodatabases whereas in future alternative delivery formats (e.g. GML) are likely to be required. Derived multiple platform specific instances of data products will be greatly assisted using a forward engineering model driven process. Likewise the requirement to update and maintain multiple related products (an activity that has not been implemented for phase 1 products, but that will need to be addressed in the longer-term) will be greatly assisted through model driven production processes.

3.2.1. Process descriptions in DPS

Within data product specifications, lineage metadata elements (from ISO 19115) are used to describe the process and sources used to create products. This mechanism is severely limited as it is focused on metadata as data set, data set series. It does have a mechanisms that enables users to define scopes for metadata statements that relate to specific aspects of a data set but in practice this is inadequate to deal with the feature type and feature instance level metadata needed to described how a product was generated. Furthermore production processes are complex spanning organisations, computing environments and applications. Thus it is not feasible to capture this lineage metadata (or provenance information) manually and almost impossible to adequately describe in textual form in any kind of meaningful way. Finally, users have different uses for data products and thus to determine fitness for use in a particular context, different granularity and types of lineage metadata will be required. The high level description of the process in the SH_network product may in fact provide sufficient information about process and data sources for the product to determine fitness for use. However, for the majority of sophisticated geospatial and hydrological modelling users this is likely to be insufficient.

3.2.2. Modelling and model mapping

The following provides a description of the key model and modelling mapping issues that have been identified by documenting the production process and the current and potential future role of model driven approaches.

There is still a large amount of contextual knowledge and modelling skill, and a not inconsiderable amount of effort required to re-factor models as part of the reverse engineering process. The principles for this process are starting to emerge and the growing suite of tools to handle individual steps could potentially be synthesised into a simplified and complete process with a higher degree of automation. There are a number of issues associated with levels of abstraction reverse engineered models (of databases) and the more abstract and platform independent models. The “persistence layer” or database schema tend to be organised around storage delivery optimisation whereas the abstract platform independent models (PIM) are ‘normalised’ to minimise ambiguity and inconsistencies and highlight key semantic relationships. These different abstractions reflect a different rationale and design philosophy. However there is a need to understand how these models need to be mapped to each other.

The mapping analysed in this report, is based on simple automated name matching, but can be manually augmented. Currently, tools for mapping are fairly crude and will need to enhanced based on a deeper understanding of requirements for mapping. Manual intervention required to map between models tends to require review of documentation and actual data – capturing this would appear to have benefits in bringing key information into a predictable locations. These activities need to described in methodology documents necessary to operationalise and sustain mapping activities. However, currently, there are no real drivers to do this as tools to read and implement the mappings are not available beyond limited transformation capabilities and thus requirements for this are not yet sufficiently understood.

Common geofabric maintenance attributes are missing from the maintenance feature classes – these need to be formally linked to the system metadata model, and some form of transformation for this aspect may also need to be defined. A methodology to extract out system metadata needs to be developed so that this can be done in a consistent manner.

Several issues related to ESRI platform specific models have been identified, related to table handling and ESRI feature sub-typing. The ESRI model provides for tables of attributes that may be joined to features (an arbitrary split between spatial and non-spatial data types). In GIS use, tables are a critical component of the way information is organised and used and thus handling tables in
UML models is an important issue. A standardised way of handling this needs to be defined and automated. ESRI has a very specific model of sub-typing, using a “soft-typing” mechanism – an attribute that has a value to flag the sub-type. This works for one level of sub-typing, as an attribute can only have a single value. This leads to a discrepancy in the documentation which refers to subtypes as shown below, and the formalised mapping and formalised models which refer to generalised class AHGFWaterbody and its specialised classes AHGFReservoir:

- AHGFWaterbody.Reservoir
- AHGFWaterbody.Lake
- AHGFWaterbody.Swamp

Thus, documentation (product guide, feature catalogue and models) is slightly out of sync with some referring to feature types and the other to sub-types. A more formal and automated process of extracting documentation out of a single model will help keep products and documentation self-consistent.
4. CONCLUSION AND FUTURE DIRECTIONS

Solid Ground toolset and registry have been developed to support the evolution of information models through registering elements of these models and the business rules and controlled vocabularies that support them. Through these tools, it is possible to ensure that disparate products remain relatable through managing the relationships between their information models. The Geofabric will increasingly face these challenges. In the short term, the Geofabric and the modelling tools are co-evolving, leaving little opportunity to directly apply the tools. This situation was envisaged in the original proposal for a model-driven Geofabric design methodology, and there has been sufficient progress in the tools to suggest that a model driven approach to design is feasible and valuable. This progress has already led to external adoption of some of the basic tools, with interest in the complete approach.

In the longer term, the goals of the research are be able to automate data production, through the exploitation of information models. In order to do this it is necessary to understand how to describe derivation processes in model mappings. This will enable end users create recipes selecting features and processes to be applied to them to create derived products from the maintenance environment through the exploitation of models.

4.1. Towards model driven product development

As indicated above, schemata for data products were generated manually and checked for conformance with the conceptual model. Thus the manual schema generation processes are based on, but not driven by models. The development of Solid Ground functionality to forward engineer (model driven engineering) of product models is a longer term target for the Solid Ground.

However, currently models are not used to drive production processes and thus at this stage, models represent informative metadata. As shown in Figure 6, there is a disconnect between the model mapping that can be used to drive product production and the actual production process. It is the underlying objective of the model driven approach to support and drive aspects of geospatial information production using models.

In the context of the SWIM project, a portion of the effort to enable forward and reverse engineering of geospatial product models, has been taken on by Sparx Systems. Sparx is developing a set of capabilities within Enterprise Architect, delivered through MDG technology, to reverse engineer UML models from ESRI geodatabases using workspace files and forward engineer geodatabase schemas from UML models.

However there is still a need to undertake further research to explore the role of mapping in geospatial information production processes. Specifically:

- understand the Bureau’s requirements for and scope of models and mapping use to drive operational geospatial production;
- develop methodologies, and how to guides that enable consistent, manageable approaches to modelling and model mapping. In particular including:
  - approaches to refactoring and extracting/abstracting out platform specific and other implementation specific details
- how to reference processes in the mappings that transform feature types, for example a line generalisation applied as part of the transformation in a mapping from a cartographic stream segment feature class in the maintenance model to a generalised representation of the line in a derived product.

4.2. Models as operational metadata

In addition to the use of models to drive production, the resultant models and mappings between them, represent critical system metadata, which will be used at run-time as ‘operational metadata’. The clear articulation of use cases for exploiting models and their mappings, will greatly inform the scope of and approaches to mapping models and model elements to each other. In SWIM project 2011/12 activity 2 will focus on articulating a registry exploitation use case and building prototype functionality to enable users to browse items registered in the Solid Ground registry. The use case currently under consideration, is the development of a feature type catalogue to support data exploration and discovery. By exposing required model elements and the relationship between,
users will be able to browse and explore related feature concepts, feature types, data products and vocabularies as an integral part of data discovery process.

4.3. Towards automated production process

Critical business rules implicit in the geofabric production process, need to be handled in a manner which enables the rules to be reviewed by hydrologists, modified and rerun to explore the effect on products. In addition a coherent suite of interrelated products needs to be maintained while input data (particularly monitoring points being loaded into AWRIS) is constantly updated. Delta monitoring points will need to processed, stitched into the geofabric network and used to update products. To assist in managing and automating aspects of this production process, the SWIM project in 2011/12 will be using the Hydrologists Workbench (HWB). HWB will be used to provide a mechanism for the generation and evolution of Geofabric products. The processing algorithms and business rules developed by Bureau Geofabric team will be ported to HWB for execution. The use of HWB will enable continued experimentation to refine existing data products, development of new products and research leading to automation of product updates.

4.4. Provenance – documenting production processes

To date significant effort has also been expended in the documenting supply product production processes and emerging geofabrication processes. Documenting the logic and processes involved in the production of supply products has been undertaken in order to understand the provenance of features in the supply products and some of the design decisions taken that have influence on the nature of and relationship between features downstream in the processing chain.

More recently, to support phase 2 production efforts have focused on designing algorithms and business rules to create phase 2 products. A series of UML activity diagrams, describing these processing steps (together with the are being produced and these are available here\textsuperscript{4}: http://iwis.csiro.au/SecureDocs/AHGF_BusinessAnalysisModel/index.htm?goto=1:3:1:130. These diagrams are, in effect, manually produced graphical provenance metadata. The diagrams articulate the complexity of the production process. However, provenance research in SWIM 2011/12 will lead to automation of provenance information of this nature for phase 2 production processes.

The extraction, storage and use of provenance information for data products produced in HWB will also be explored. This will entail identification and extraction of metadata related to processes and data sources from the service layer of HWB, making this information available to provenance persistent stores.

The use of HWB and the mining of workflow metadata to create provenance information about data products, builds towards the long-term research agenda of executable data product specifications. In order to be able to automate data production is necessary to understand how to describe data inputs to processes and process on the data.

\textsuperscript{4} This activity is ongoing and models at detailed levels (particularly for the currently evolving model link network production workflow) are yet to be developed.
REFERENCES


