Future Internet **2009**, *1*, 28-46; doi:10.3390/fi1010028

OPEN ACCESS

future internet

ISSN 1999-5903

www.mdpi.com/journal/futureinternet

Article

A Method for Automating Geospatial Dataset Metadata

James K. Batcheller *, Bruce M. Gittings and Robert I. Dunfey

Institute of Geography, School of GeoSciences, University of Edinburgh, Drummond St., Edinburgh, EH8 9XP, UK; E-Mails: bruce@ed.ac.uk (B.M.G.); robert.dunfey@shell.com (R.I.D.)

* Author to whom correspondence should be addressed. E-Mail: jk.batch@ed.ac.uk.

Received: 28 August 2009; in revised form: 1 November 2009 / Accepted: 6 November 2009 / Published: 10 November 2009

Abstract: Metadata have long been recognised as crucial to geospatial asset management and discovery, and yet undertaking their creation remains an unenviable task often to be avoided. This paper proposes a practical approach designed to address such concerns, decomposing various data creation, management, update and documentation process steps that are subsequently leveraged to contribute towards metadata record completion. Using a customised utility embedded within a common GIS application, metadata elements are computationally derived from an imposed feature metadata standard, dataset geometry, an integrated storage protocol and pre-prepared content, and instantiated within a common geospatial discovery convention. Yielding 27 out of a 32 total metadata elements (or 15 out of 17 mandatory elements) the approach demonstrably lessens the burden of metadata authorship. It also encourages improved geospatial asset management whilst outlining core requisites for developing a more open metadata strategy not bound to any particular application domain.

Keywords: geospatial metadata; metadata automation; data documentation; discovery metadata; feature metadata

1. Introduction

Metadata have been key to the development of geospatial data sharing initiatives since their first emergence. From early mention of ‘data registers’ of spatial assets in the Chorley Report [1], through pioneering initiatives such as the Federal Geographic Data Committee’s (FGDC[1]) National Geospatial

[1] http://www.fgdc.gov/

Data Clearinghouse in the US and the National Geospatial Data Framework (NGDF) in the UK, to more recent Internet services such as the *Geospatial One-Stop*, *Gigateway* and the latter's academic counterpart *Go-Geo!*, what has become known as geospatial resource metadata has consistently assumed a central pillar around which efforts have evolved [2-7]. Predicated on the basis that textual metadata records are more widely accessible to location and retrieval techniques than the geospatial holdings they describe, such services aim to facilitate and promote data exchange by registering such data surrogates within searchable catalogues or clearinghouses. Users poll catalogue contents based on dataset properties depicted in the metadata (e.g., spatial coverage, theme, keyword); result sets inform decisions as to whether datasets should be pursued, and where they may be ultimately located.

While it could be argued that the core function of metadata in driving data sharing efforts has remained largely unchanged from pioneering implementations to date, encoding strategies and the mechanisms through which metadata records communicate dataset availability have seen significant progression. Fuelled principally by a desire for interoperability, domain specific metadata strategies have made way for broader standards at regional and national levels, culminating in the recent convergence around International Organization for Standardization (ISO) conventions [8-10]. Similarly, computational frameworks used to enable geospatial resource discovery continue to witness a shift away from applications developed primarily for other disciplines, such as the library community's Z39.50 protocol [6,11,12] to more recent strategies designed specifically for the geospatial domain, in particular those based upon Open Geospatial Consortium (OGC) specifications. Sophisticated proprietary offerings in addition to those produced by the open source community have arisen, providing geo-centric solutions with the potential to empower geospatial data providers wishing to publicise their holdings while also facilitating end-user discovery [13-17].

Further a-field, technologies with the potential for enhancing geospatial data exchange continue to surface. Improved networking infrastructures, coupled with the heightened availability and plummeting costs of high-speed broadband and wireless connectivity have in turn served to streamline the access and transfer of the high data volumes often characteristic of geographic datasets [18-20]. The appearance of Geography Mark-up Language (GML) has helped alleviate many of the concerns relating to data compatibility and interoperability, providing an open dialect for data transfer not bound to specific software offerings [21-23]. Meanwhile, maturing open web portrayal techniques such as those propounded by the OGC offer scope for integration with current discovery services, allowing spatial data to be previewed prior to committing resources for procurement [24,25]. Similarly, emerging geoprocessing services such as those aligned with the OGC's Web Processing Service specification offer the potential for remote processing prior to data retrieval, maximising transfer efficiencies whilst minimising end-user software and hardware demands [26-29].

And yet, considering the role that geospatial metadata retains in facilitating the discovery and exploitation of the very data resources on which the many of the aforementioned services depend, the lack of practical efforts aimed at providing easier methods for its generation is telling. While the OGC rightly focuses on issues of abstract interoperability and interface specifications [30], actual implementations of metadata-based specifications are invariably predicated on the assumption that metadata instances will be forthcoming (e.g. United Nations Food and Agriculture Organisation's GeoNetwork). For initiatives that explicitly address metadata generation using browser and desktop-based metadata editors (e.g., Gigateway, Go-Geo!), assistance rarely exceeds guiding manual metadata

completion and consequently does little to surmount perceptions of a task as being tedious and of low priority [31,32]. Some proprietary packages (e.g., ESRI's ArcGIS, Cadcorp's SIS) do incorporate metadata management applications capable of generating rudimentary element sets for given geospatial resources, although these invariably demand significant human input to meet compliance with even the most concise of standards.

It is this need for human input that contributes towards a 'metadata bottleneck' [33] that hinders documentation, whether for resource discovery, management or exploitation. Attempts to address this impedance have to date largely fallen within two camps. Those of the first strive to motivate and mobilise metadata authors (education drives and workshop events) [4,12,34]; those of the second assume mobilisation, and focus on steering metadata completion (metadata profiles, guidelines and editors). While the role both strategies play in encouraging data documenting practices can be significant, their impact in diminishing the burden of metadata authorship remains open to question.

This paper proposes a method to mitigate this burden through the automation of metadata authoring process steps. Whether from the perspective of cataloguing internal digital holdings, exposing marketable datasets online via metadata services or fulfilling legal requirements by contributing to national spatial data infrastructures (NSDI), the implications of metadata automation are several. By providing computational support to metadata authors, costs of generation can be lessened whilst rendering the prospect of creation less daunting for those yet to undertake it. Such support also reduces the opportunity for human error during metadata record completion; it enables more records to be produced with equivalent effort; and it facilitates easier update of existing metadata following changes to their underlying data. Further, authoring resources can as a result be released for application elsewhere, whether for more intellectually challenging documentation tasks (e.g. quality control, descriptive metadata) or otherwise. It is thus seen to play an important role in efforts to diminish the stigma associated with metadata and its creation, a critical consideration if documentation practices are to be more widely promoted.

2. Study approach

One method for automating metadata production is grounded in the premise that a dataset, its contents and ambient computing environment can be leveraged for their contribution. In Batcheller [35], the approach presented focussed around a proprietary GIS and standard-based data management protocol. Metadata elements were computed from a dataset's position within a folder hierarchy, the dataset construct itself, as well as from sources pre-compiled both manually and by the host GIS. Metadata were also extracted from dataset data, but here the approach was limited in two notable aspects – attribute data did not adhere to a formal schema and consequently did not reflect conditions typical of production environments; and no effort was made to mine the geometric component distinctive of geospatial datasets.

In the current paper we present an extension to the original approach, one that expands the contribution of both spatial and aspatial data in generating dataset metadata automatically. We propose that the exertions of manual data documentation can be minimised by assuming a holistic approach to distinct processes such as data modelling, creation, storage and management; the founding work is briefly outlined in accordance with this view. We go on to elaborate an extensible, interoperable

schema for structuring and documenting attribute data at the sub-dataset level and detail its novel use in generating higher-order metadata. We introduce the concept of the boundary reference layer, illustrating how it computes metadata items from dataset geometry beyond the conventional provision of bounding coordinates and how the process may mitigate element-specific update latency. Further, we describe the use of published guidelines to support completion of descriptive metadata, the computational creation of which is frequently overlooked due to its perception as an exclusively manual endeavour. Finally, we describe a means for incorporating geospatial asset visualisation for the purpose of decision support.

2.1. Software environment

With the assumption that geospatial resource documentation commences in close proximity to the datasets being depicted, the proposed utility was developed for ESRI's ArcCatalog, the data and metadata management component of its ArcGIS 9.1 suite. Incorporating an authoring tool within the same application used for dataset registration and maintenance carries with it a number of advantages: it minimises the data-metadata disconnect at source, alleviating latency concerns; editing and update practices for both data and metadata are consolidated within a single package, enabling workflow integration; and it negates the need for external editors where authoring effort may be duplicated and which presents further opportunity for introducing errors.

The application logic that drives metadata generation falls into one of two categories, as described by Greenberg [36]. Routines that *harvest* gather items already held as metadata within the given domain, yet often dispersed throughout it; those that *extract* must transform data read from resources within the domain into metadata-ready form. Metadata elements are consequently computed from three principal sources – the dataset requiring documentation, its immediate computing environment and pre-prepared content unlikely to vary between datasets of the same origin.

2.2. Geospatial metadata standard

The UK GEMINI (GEo-spatial Metadata INteroperability Initiative, a.k.a. GEMINI) convention was selected to instantiate the metadata output of the prototype. As the prevailing discovery standard in the UK it offers a widely recognised, succinct format with a proven pedigree in geospatial production environments, one in which the performance of the present work can be demonstrated. Derived in accordance with *ISO 19106 Geographic Information: Profiles* from both *ISO 19115 Geographic Information: Metadata* and the UK eGovernment Metadata Standard (*eGMS*), GEMINI outlines thirty-two elements (Table 1) that are readily mapped to other ISO-based formats [37].

Table 1. The UK GEMINI 32 metadata element set - optional elements in italics.

Element	Description
Title	Dataset name
Alternative title	*Alternative name*
Dataset language	Language used
Abstract	Narrative summary describing the dataset
Topic category	Main themes of dataset (high-level categories)
Subject	Topic of the dataset content (low-level categories)

Table 1. *Cont*.

Element	Description
Date	Data capture period
Dataset reference date	Date of dataset publication
Originator	*Originating person or organisation*
Lineage	*Dataset pedigree*
West bounding coordinate	Western limit of dataset
East bounding coordinate	Eastern limit of dataset
North bounding coordinate	Northern limit of dataset
South bounding coordinate	Southern limit of dataset
Extent	Geographic identifier of dataset
Vertical extent information	*Vertical domain of dataset*
Spatial reference system	Name of spatial reference system
Spatial resolution	*Capture precision of data*
Spatial representation type	*Method of spatial representation*
Presentation type	*Method of data manifestation*
Data format	Digital format of data
Supply media	*Method of data supply*
Distributor	Distributing organisation
Frequency of update	Prescribed frequency of data update
Access constraint	*Rights of data access*
Use constraints	*Rights of data use*
Additional information source	*Source of further details about dataset*
Online resource	*Online sources of dataset*
Browse graphic	*Illustrative sample of dataset*
Date of update of metadata	Last date of metadata update
Metadata standard name	*Name of metadata standard and profile used*
Metadata standard version	*Metadata version used*

3. Founding work

3.1. Preliminary generation – dataset initialisation

Metadata generation commences upon registration of new data within the ArcCatalog application. Properties set by the user during dataset initialisation (e.g., *Spatial reference system*, *Title*, *Bounding coordinates*) as well as those taken from the software's default configuration (e.g., *Dataset language*) are collected by the package and stored alongside the dataset as an XML-based metadata record that is available for subsequent editing, or collection as in the current scenario. Other existent dataset properties not specifically treated as items of metadata by ArcCatalog (e.g., *Alternative title*, *Vertical extent information*) are not accessible in this way and must therefore be targeted by extraction routines run against the dataset.

3.2. Geospatial data management protocol

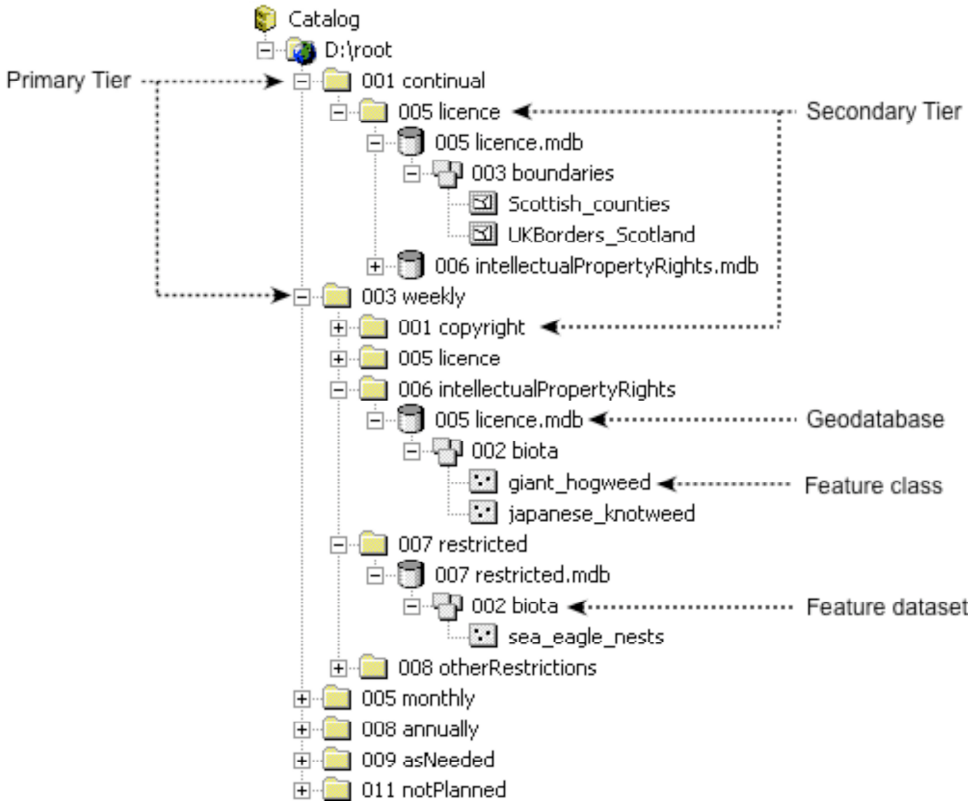
Participating datasets are organised within a multi-tiered folder hierarchy, the members of which are tagged on the basis of the data to be held therein; tag nomenclature is based upon domain instances of

entities as defined by the GEMINI standard. Here a two-tiered hierarchy is used to store personal geodatabases, the defining entities of which were chosen for their presumed suitability for categorising and filing datasets (Table 2). Data are further classified within each geodatabase using the ArcGIS feature dataset and feature class constructs, the latter representing the geographic layer to be documented. What results is a virtual, predictable path to the data that not only enables dataset retrieval but also serves to contribute a subset of dataset properties for documentation purposes (Figure 1). Consequent population routines manipulate this path and assign its components to the appropriate metadata fields. The entities used and their domains are detailed in Table 2; a prototype hierarchy is provided in Figure 1.

Table 2. Data container definitions based upon those UK GEMINI entities deemed appropriate for dataset categorisation. Actual container instances are labelled using the respective UK GEMINI domain values.

Data container	UK GEMINI Entity	UK GEMINI Domain
Primary tier	Frequency of update	MD_MaintenanceFrequencyCode
Secondary tier	Access constraints	MD_RestrictionCode
Personal geodatabase	Use constraints	MD_RestrictionCode
Feature dataset	Topic category	MD_TopicCategory
Feature class	Title	Free text

Figure 1. A prototype data container hierarchy, yielding five metadata elements (adapted from Batcheller [35]). Containers are instantiated on arrival of new datasets; an entire UK GEMINI-based hierarchy need not be constructed.



3.3. User-defined content

Geospatial metadata records arising from a single source invariably contain information that either remains constant or changes infrequently across instances, such as the contact details of the data producer. For elements that differ according to user or location (such as *Distributor* details where multiple distribution sites exist), information stored in content templates are loaded on the basis of local operating environment variables such as username or domain. For organisations participating in geospatial data sharing initiatives, default domain values for GEMINI elements such as *Presentation type* and *Supply media* may be set to *mapDigital* and *onLine* respectively – entries that can be hard-coded into any automated approach but which can be overwritten if needed. *Metadata standard name* and *Metadata standard version* elements are similarly be initialised, while event related entries such as *Dataset reference date* and *Date of update of metadata* can be automatically time-stamped.

4. Expansion of previous work

4.1. Dataset derived metadata – feature metadata

Attribute data have the potential for considerable contribution towards automated documentation given their role in explicating the discrete features that constitute a dataset; tapping this potential is nevertheless complicated by the variety of ways in which attributes may be ordered and encoded. One option is to index commonly occurring values from those semi-structured attribute fields not conforming to specific data standards; as the data are encoded in an inconsistent fashion however, any extraction routine run will necessitate human mediation of each candidate metadata value returned. Spatial data standards impose a predictable structure on attribute schema and can thus address this need for constant review, but their diversity complicates selection even prior to contemplating interoperability-related concerns [38] and the bloat their verbosity can add to data preparation and upkeep.

A more concise alternative for enforcing predictable attribute structure is proposed, one that allows for the association of a metadata schema with the features of each dataset – and in turn, address concerns articulated by Hunter [39] and Devillers *et al.* [40] relating to the depiction of geometric primitives. Accordingly, the approach allows for the consistent referencing of member fields whilst providing a mechanism for tracking a dataset's atomic features. Distinct from conventional theme-driven standards, the proposed feature metadata approach was designed to be:

- domain-independent, compatible with existing standard-based techniques and yet easily extensible to permit flexible adoption
- associated specifically with the geometric component of a dataset, thereby disengaging feature metadata from potentially inter-changeable theme-based data and enabling its persistence
- straightforward to apply to existing data models, with potential for automated completion at the point of survey of new geographic features
- capable of contributing towards automating metadata creation at the dataset level without complicating data creation and maintenance

While a review of existing (published) feature metadata strategies did present a number of pre-existing alternatives (most notably the State of Maine's GIS Feature Metadata Recommendation 2000[2]), none met all of the above conditions. As a consequence the current paper elaborates a course more closely aligned with ISO-based practices, heeding the stated developments in the standards community.

ISO 19115:2005 Geographic Information – Metadata [41] for instance details a content schema for the documentation of geospatial data, but its focus lies with the depiction of geospatial resources at the dataset, and to a lesser degree, dataset series level. Metadata at the sub-dataset level are presented within a metadata hierarchy, but suggested implementations only include definitions at these levels when exceptions occur. Further, as metadata conforming to ISO 19115 are held discrete from the resources they describe by convention – regardless of its granularity – this treatment for metadata is on its own insufficient when reviewed for the current purpose.

The ISO 19109:2005 Geographic Information – Rules for Application Schema standard meanwhile allows for the definition of conceptual data models that define the logical structure of an application's data [42]. Geographic feature types are classified based on a structure defined by the General Feature Model (GFM); feature type definitions (detailing feature attributes, operations and association roles) may be elaborated in feature catalogues. Of particular interest is its specific treatment for feature attributes as well as the general ability to integrate any ISO 19109 application schema with other ISO standard schemas. Here, any feature attribute (GF_AttributeType) can have atomic metadata items associated with it by sub-classing entities beneath the GF_QualityAttributeType specialisation of the GF_MetadataAttributeType entity (Figure 2). Attribute types accordingly defined (specifically, to carry feature metadata information such as quality) obtain their value type definitions and value domains from the ISO 19115 MD_Metadata entity.

A six-field element set similar to what can be expected within formal production environments for the purposes of feature-level data tracking was consequently elaborated and mapped using the ISO framework (Table 3). Aliases serve as shorthand field labels, while full element definitions are included within dataset metadata instances by default to permit interpretation. Field domains may be enforced in applications where stringent compliance is required, however these were not currently applied. For new data holdings, the schema may be incorporated at the point of survey through the inclusion of a data dictionary within the surveying equipment, thereby facilitating pre-population prior to registering the dataset with a GIS. Existing datasets may have equivalent entries mapped to the schema if such entries exist; alternatively the schema may be appended to an incumbent attribute table for subsequent completion. Once implemented, fields may be analysed and extracted in accordance to the dataset metadata element for which they may be applied.

The *origin* field is used to populate the GEMINI element *Originator* in the current prototype. Multiple entries are generalised to indicate the single dominant originator reflecting instances where only one entry is allowed; multiple summary values may also be used where permitted. *Spatial resolution* is calculated from *precision* using the lowest common denominating value of feature precision, although this may also be computed on the basis of average value if preferred. The date range *Date*, used to indicate the data capture period, is calculated with functions identifying the

[2] http://apollo.ogis.state.me.us/standards/flmeta/fmrecommend.htm

minimum recorded *capture_date* and maximum *edit_date* of features within the dataset. Dataset provenance, as treated by the *Lineage* element to store 'information about the events or source data used in the construction of the dataset' [37] is similarly populated. Metadata fields used to track feature history (*i.e.*, *capture_process*, *capture_date*, *editor*, *edit_date*) contribute here, used to append details of each process step to the *Lineage* element.

Figure 2. Attributes of ISO 19109 feature types, with an emphasis on the route to metadata-relevant subclasses. Feature metadata instances are catered for by GF_QualityAttributeType's dependency on the ISO 19115 entity MD_Metadata. (Subset taken from ISO 19109:2005).

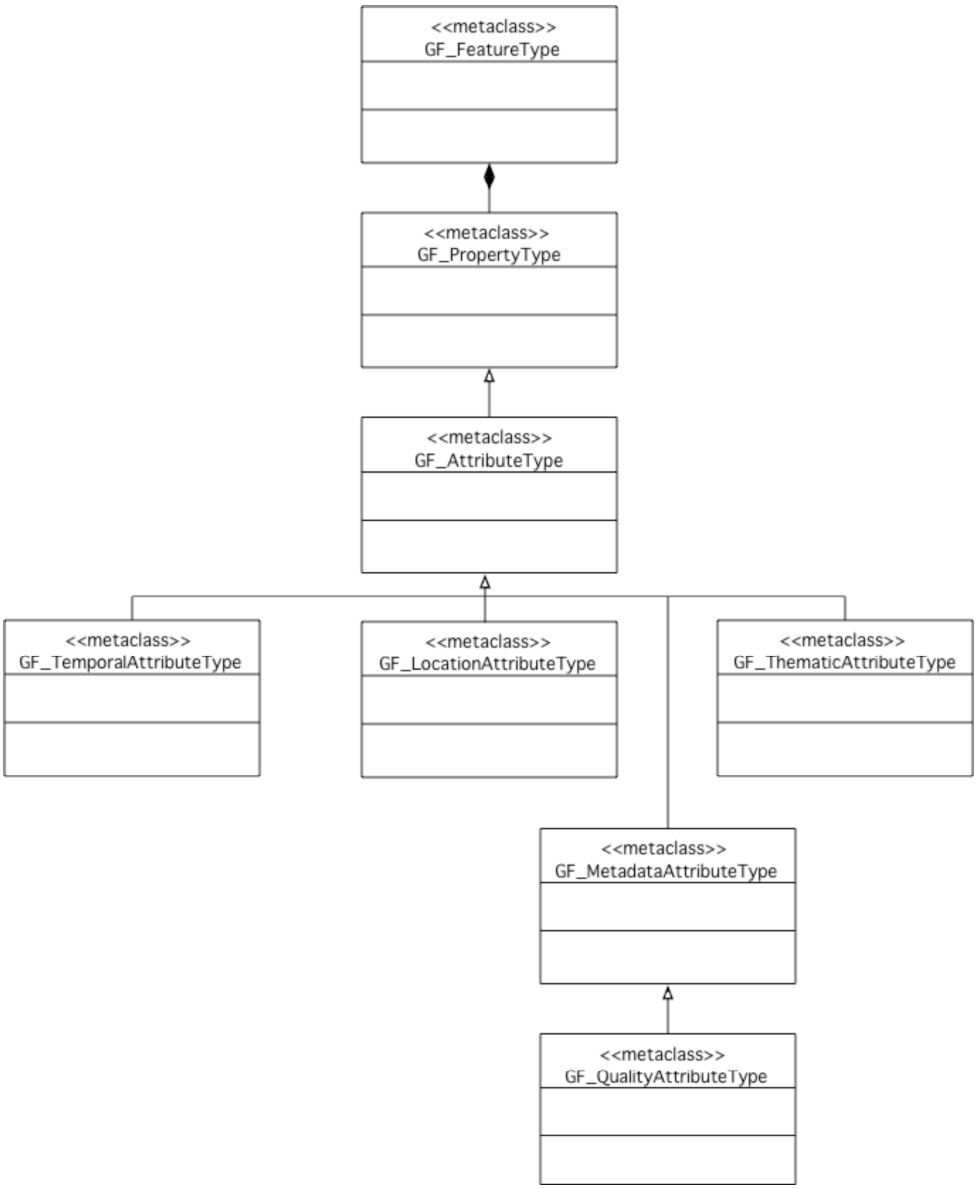


Table 3. ISO-based feature metadata schema, defined using ISO 19115 entities according to ISO 19109 guidelines (all being subclasses of MD_Metadata.dataQualityInfo). Both organisationName entries (origin, editor) may be replaced with (or supplemented by) individualName if required by the domain of application.

Field	Definition	Alias
Originating organisation	DQ_DataQuality.lineage> LI_Lineage.source> LI_Source.sourceCitation> CI_Citation.CI_ResponsibleParty.organisationName	origin
Originating Capture Process	DQ_DataQuality.lineage> LI_Lineage.sourceStep> LI_ProcessStep.description	capture_process
Originating Capture Date	DQ_DataQuality.lineage> LI_Lineage.sourceStep> LI_ProcessStep.dateTime	capture_date
Precision	DQ_DataQuality.report> DQ_Element.DQ_PositionalAccuracy.result	precision
Editing organisation	DQ_DataQuality.lineage> LI_Lineage.processStep> LI_ProcessStep.processor> CI_ResponsibleParty.organisationName	editor
Edit Date	DQ_DataQuality.lineage> LI_Lineage.processStep> LI_ProcessStep.dateTime	edit_date

4.2. Dataset derived metadata – geometry

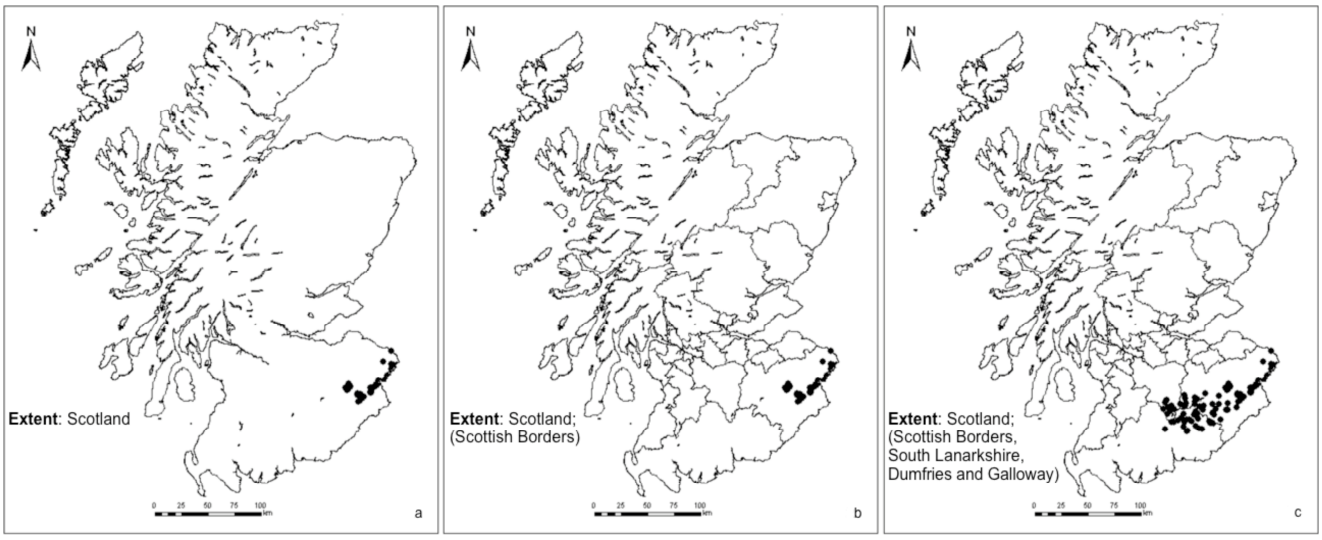
Deriving metadata from data geometry is conventionally limited to the calculation and update of projection-dependent bounding coordinates of the entire dataset. Typically employed to provide a rudimentary spatial component for searching indexed metadata catalogues, such coordinates do not however convey extent information that is easily interpreted by end-users. Boundary names (e.g. political, administrative regions) provide more user-friendly extent attribution but as these are not directly coupled with the spatial component of the dataset, any change to the data's geographic bounds will not be reflected in such attribution unless specifically detected and manually addressed.

Such issues are circumvented through the association with the data a layer depicting application-relevant boundaries, enabling the automated identification of a dataset's potentially evolving extent by a process of reverse geocoding. Feature sets are programmatically overlain with such boundary reference layers, returning the name of the enveloping region. Extent identifiers within the metadata are accordingly coupled with underlying spatial data, removing the need for manual nomination and update that can be subjective and open to error.

A related issue pertains to spatial resolution. The base GEMINI specification for instance outlines an *Extent* field domain of high level areas such as England, Wales, Great Britain, British Isles, but sole inclusion of a reference boundary layer with these regions may be insufficient for applications with more fine-grained requirements, as with those of local government. Incorporating subsequent reference layers of increasing scale can facilitate improved extent identification and help relay more precise information within the resulting metadata. To illustrate, a test dataset containing the geographic distribution of the invasive Giant Hogweed species (*Heracleum mantegazzianum*) in the south of Scotland is used (Figures 3, a-c). Figure 3a depicts the problem of registering an extent of Scotland for a rather localised phenomenon found in the south east of the country. Introducing an administration

boundary layer provides more detail (Figure 3b), showing that the recorded extent is in fact confined to the Scottish Borders. Figure 3c illustrates a hypothetical spread of the weed outwith its original confines and into South Lanarkshire and Dumfries and Galloway; employing such a reference boundary layer not only enables the automatic detection of cross-border spread at run-time, it produces richer metadata on which more informed decisions may potentially be based.

Figure 3. (a-b): Geographic extent estimation of *Heracleum mantegazzianum* in the south-west of Scotland, 2003-05, and the effects of boundary reference layer granularity (data courtesy of the Tweed Foundation). **(c):** A hypothetical spread of the species and the districts detected using the boundary reference layer approach.



4.3. Abstract seeding

Metadata conventions typically permit the inclusion of free-text resource descriptions via an *Abstract* field that stands in contrast to other domain-controlled elements. While it is perhaps unrealistic to expect metadata automation efforts to adequately complete such narrative entries, the process may nevertheless be assisted through the seeding of abstract fields with the information items from which they are comprised. Guidelines for completing UK GEMINI metadata for instance outline a checklist for abstract completion that demands 13 distinct components, many of which may be extrapolated or approximated from existing (populated) sources (Table 4). Contribution towards abstract creation is thus performed by seed routines that return abstract items into the field for subsequent elaboration manually. Serving not only to mitigate the effort required to complete the element, the approach ensures consistent inclusion of abstract-relevant items that hold the potential of being overlooked where authoring is left unassisted.

Table 4. Guidelines for completing abstract content (adapted from recommendations published by Gigateway at http://www.gigateway.org.uk/metadata/standards.html/).

Guideline	Seed source
What the dataset depicts	Alternative Title
Area of coverage	Extent
Period of coverage and frequency of update	Date, Frequency of update
Data capture scale / resolution	Spatial Resolution
Data capture method	capture_process
Suggested uses for data	Topic
Category of features depicted	Subject
Details of limitations in data	-
Data linkages	Feature ID field
Data originator(s) / editor(s)	Originator, origin, editor
Data model	Spatial representation type
Data format	Data format
Data series	Sibling feature classes

4.4. Asset Visualisation

While GEMINI was adopted to illustrate the opportunities for automating metadata regardless of its ultimate use, it is perhaps worth considering the specific case of enabling resource discovery. GEMINI elements *Online resource* and *Browse graphic* are of specific interest here; the former enables pinpointing of the resource described, the latter is designed to support decisions based on whether the resource should indeed be pursued. A common approach to address these entries has been to provide a URL for the distributor's website while including a manually generated image snapshot of a subset of the data. The alternative proposed in the current work aims not only to facilitate improved metadata completion but also to provide for more effective visualisation with the potential for fast-tacking access to the data in question.

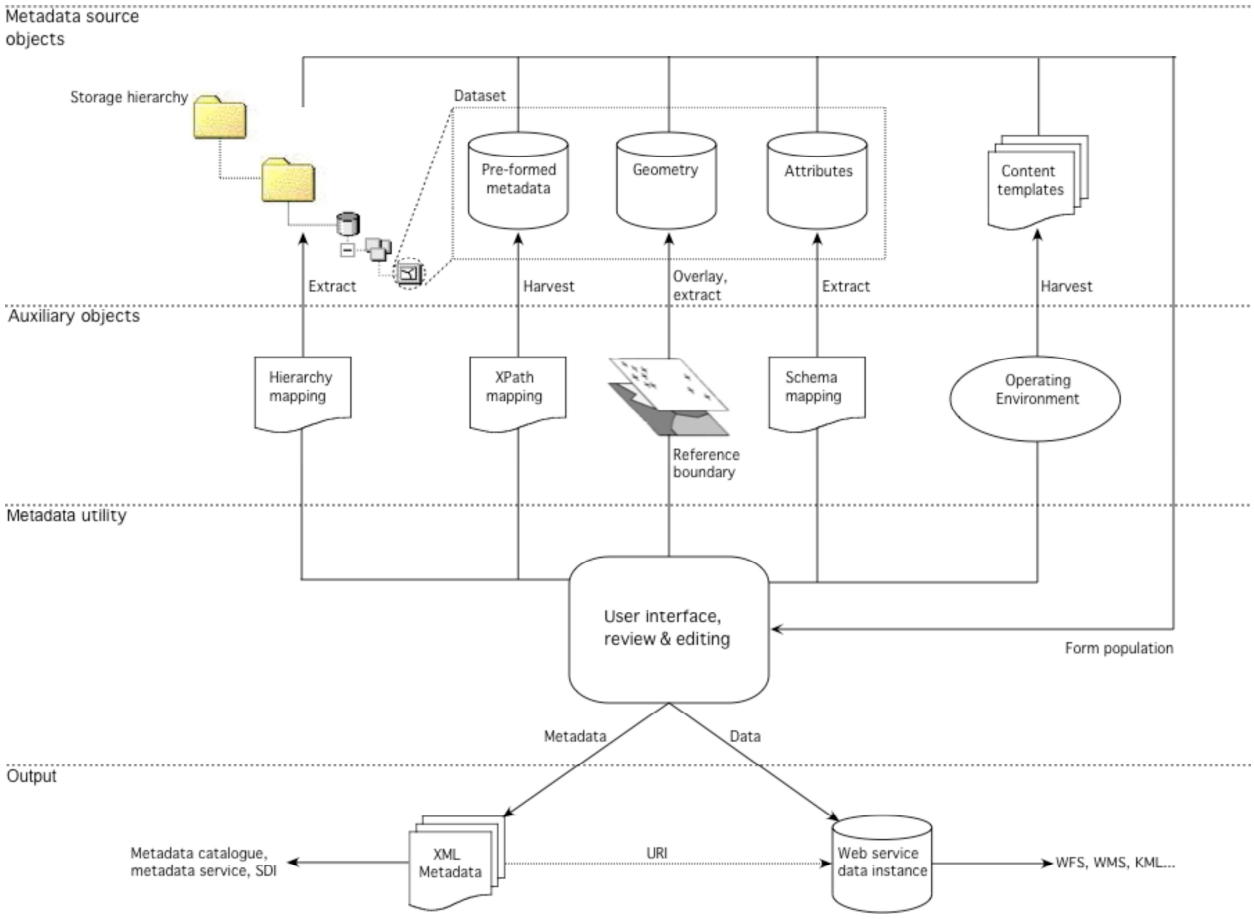
Using OGC-compliant web portrayal services, datasets can be 'broadcast' for immediate use when unlicensed, or identified for subsequent purchase. Layers free of licensing restrictions may be served in accessible Web Feature Service (WFS) format whereas proprietary holdings may be visualised in Web Mapping Service (WMS) format prior to procurement. Service-specific Uniform Resource Locators (URLs) are subsequently written to their associated metadata record, coupling it with the portrayal service.

5. Execution

The utility was coded using Microsoft's VB.NET and ESRI's ArcObjects framework, embedded within ArcCatalog as a dynamically linked library and exposed as a custom button within the application toolbar. Data instantiated as feature classes within the folder-geodatabase hierarchy are selected individually within the application's file browser and the tool is engaged. Harvesting and extraction routines, initiated from a form interface, together compute metadata elements for display within the form where they are made available for quality control and update prior to final output

(Figure 4). Pre-formed metadata items are harvested from ArcCatalog's XML store using XPath expressions that allow referencing of specific elements. These expressions are read from a rudimentary crosswalk file that not only details the individual element sources within the application's XML metadata repository but also describe (in XPath) where form values will be written to within the XML-formatted GEMINI output document. Initial population routines return all elements in the event of prior completion; further routines may be executed selectively where elements are incomplete or require revision.

Figure 4. An overview of approach's routines, the metadata sources on which they operated (assisted by its associated auxiliary object). Computed values are displayed on a form interface where they may be modified by hand prior to validation and eventual output.



Elements are computed from feature metadata content at run-time, referenced via their corresponding field names; routines index, summarise and identify maxima and minima according to the GEMINI element they contribute to. Dataset geometry is overlain with the incorporated reference layers to generate boundary names that are then added to the *Extent* field and later used to seed the *Abstract* entry. The application extracts elements from the storage hierarchy by parsing the dataset path and assigning its constituent tokens to form fields according to a path – element lookup table. Pre-authored content templates held within the file system as XML are harvested on the basis of active username, the contents of which are again addressed using XPath. Seeding of abstract content is

performed as a penultimate event once other fields have been populated; contributing elements are harvested directly from the form and passed to the abstract field where they are elaborated upon manually. Metadata records are output as XML files once they have been evaluated and timestamps to actualise the *Date of Metadata update* element have been applied. Metadata passing validation may now be exploited as surrogates for the data they depict, whether locally or following contribution to an off-site repository.

Providing support for visualisation may be implemented by combining the metadata output operation with the conversion of the active dataset to an intermediary web portrayal format. Datasets are exported to an Apache Tomcat web container from where they are served as WFS layers by the open source Geoserver application; WMS and Keyhole Mark-up Language (KML) formats may also be configured for visualisation according to the terms of access for the associated data. Following initial service composition, asset-specific URLs are written to their corresponding metadata records; subsequent data updates overwrite existing data serving instances, forgoing the need for further link revision.

6. Metadata output

Of the seventeen mandatory elements detailed by the UK GEMINI convention, fifteen are automatically populated, the exceptions being the *Abstract* (partial) and *Dataset reference date*, the latter indicating a notional publication date. Twelve further optional elements are generated (two arising from default values hard-coded into the utility) leaving three entries requiring manual completion – these being the elements associated with supplemental dataset description and portrayal: *Online resource*, *Browse graphic* and *Additional information source*. While a total contribution of twenty-seven entries will undoubtedly mitigate the burden of authorship, it should be noted that there is little treatment for compound elements in the current approach. Exactly half of the defined thirty-two-element set defined by GEMINI permits multiple occurrences; little direct attempt has been made to address these (*Topic category*, *Subject* and *Extent* aside). A number of compound elements can be catered for using supplemental content templates and defaults as needed; entities otherwise populated will most likely necessitate manual intervention.

The most apparent advantage of the approach is the number of metadata items produced, although further benefits may also be seen. For one, metadata content is more dependent upon the data depicted, increasing the likelihood that dataset characteristics are more accurately reflected in the surrogates produced. The process that facilitates this admittedly only displaces rather than eliminates the burden of effort from data documentation to data management, and could thus be construed as representing a false economy. However, the contention is that the increased emphasis on dataset configuration and categorisation practices will invariably have positive consequences for data asset quality and management, while simultaneously supporting eventual metadata output.

7. Discussion and Conclusions

What has been presented can be viewed as a holistic approach to metadata automation. Various data and metadata creation, update and management process steps have been decomposed and subsequently employed to contribute towards the automated generation of geospatial metadata records. Standardised

data encoding and storage practices have been extended beyond their traditional use in facilitating data upkeep, location and query to provide tangible support to metadata authoring. Entities drawn from both a dataset's enveloping computing environment and its internal constituents contribute towards documentation efforts through a system of annotation that permits such objects to be consistently referenced via a custom utility. What remains is a limited number of elements that require manual processing prior to human mediated quality control.

There nevertheless remains scope to further extend the current work. Demands for incorporating the socio-political context in which datasets are created and curated (as suggested by Comber *et al.* [43]) could be met by harvesting or linking to 'project-level' metadata such as that maintained within the Scottish Executive's GI Projects Index Registry[3] (part of their 'One Scotland – One Geography' initiative). Treatment for experiential metadata suggested by the same study could further be accommodated via the inclusion of voice recognition technology within the host GIS to enable easy capture of user-centric data perspectives while it is being manipulated and in turn enrich the outcome of documentation. Similar provision could be included within surveying equipment to permit registering of erstwhile neglected information such as the environmental conditions at the instance of data capture, providing for more detailed data quality assessment at the feature level if later required.

Complete abstract entries could be approximated through the use of pre-composed sentence templates into which seed values may be inserted; whether this would act to discourage human input or complicate validation should be considered. Instantiating a reference boundary dataset library as an accessible Application Programming Interface (API) -driven web service meanwhile would avoid the need to bundle such layers within a custom utility and offer a single point of management when updates to boundary data are required. Further scope for exploiting dataset geometry may also exist through the development and application of pattern matching approaches; a catalogue of distinctive geographic features for instance could be used to extrapolate spatial reference details of data assets lacking such information through the cross-referencing of geographic footprints with those of known aspect. Potential for improvement may also exist where the underlying metadata standard is concerned. The *Frequency of update field* for instance is employed to indicate the prescribed data maintenance period; dynamically binding this element with data update events would provide a more accurate reflection of management history, rather than an arbitrary frequency estimate that may or may not be adhered to.

We would offer that the value of implementing a feature metadata schema goes farther than providing a more sophisticated means of data tracking. Indeed, the advent of feature-centric strategies such as multi-user corporate databases and Internet-enabled delivery mechanisms like WFS has witnessed an increase in the ability to remotely access the atomic components of data resources, rendering the need to retrieve host datasets in their entirety unnecessary. Associated efficiency gains (*i.e.*, transport, processing) are nevertheless tempered by an uncertainty surrounding the pedigree of data thus retrieved, as conventional feature delivery approaches inconsistently depict the metadata of source collections. In contrast, the difficulty of collating resource documentation when deriving data from multiple-sourced feature sets illustrates the problem when dataset metadata is indeed accommodated. Implementing a metadata strategy focussed on the feature level may well be warranted

[3] http://www.gisprojects.net/

in either case: whether to complement existing dataset metadata with more detailed information on individual features, or to contribute towards the automated generation of metadata subsequent to deriving new datasets from disparately sourced feature aggregations.

As is often the case with software deployment, the contribution of a particular solution will tend to be maximised when tailored to the local context, the consequent trade-off being solution portability. While the current work is presented within a specific context in terms of data configuration, metadata output and software application, it nevertheless alludes to the requirements a more generic approach not bound to specific domains or computing environments would need to satisfy. First, access to geospatial data stores should be transparent and not reliant upon the presence of specific, third party software. The personal geodatabase format adopted above was chosen so as to restrict the degrees of freedom of the current analysis and also due to it being a commonly used format positioned between single-user file-based hybrid storage strategies and multi-user, integrated stores. It is however a closed proprietary format and thus necessitates the use of its host application to provide full access to its constituents. Initiatives such as the open source Feature Data Objects (FDO) project[4] could provide significant assistance in bypassing such restrictions, providing an extensible API for "manipulating, defining, and analyzing geospatial information regardless of where it is stored." Second, attribute data should be amenable to analysis and mining. Access to the geometric content of datasets is to a degree predictable across different formats (cf. return bounding box, get projection); consistent polling of aspatial data on the other hand is approximated through the imposition of standardised schema. Of the three strategies mentioned (keyword index of semi-structured data; formal spatial data standards; feature metadata), implementing an interoperable feature metadata approach may present the most promise, particularly in scenarios of deriving data from multiple sources as mentioned earlier.

Third, eventual metadata output should not be restricted to a particular standard or profile. The choice of convention has direct bearing on what elements are automatically generated and the format in which they are output. The emergence of ISO-based standards as the dominant initiative within the geospatial community arguably makes the task of catering for multiple output formats less problematic, permitting the elaboration of a base specification from which custom profiles may be derived. And finally, encapsulating a generic approach within a platform independent solution will maximise adoption and avoid marginalisation of any single user community. Existing initiative-driven metadata editors, whether browser-based or developed using a cross-platform software development kit (SDK) such as Sun Microsystem's Java SDK provide a sound basis on which more comprehensive metadata management approaches unhindered by operating environment-related restrictions can be extrapolated.

In the end, conventional metadata creation is unlikely to ever overcome its perception as an inconvenience, no matter how intensively its benefits are espoused. It therefore behoves the proponents of metadata practices to find ways of mitigating the burden of authorship, rather than solely pursuing the traditional dual-pronged 'carrot and stick' strategies that currently pervade. Regardless of whether a high-impact, customised approach is taken or whether a generic solution is developed for broader consumption, a niche for both certainly exists. Conditions such as the volume of available resources,

[4] http://fdo.osgeo.org/. It should be noted that the personal geodatabase remains inaccessible via this and other non-vendor APIs.

incumbent computing infrastructure, in-house expertise and perhaps most pertinently, the extensiveness of geospatial data holdings will all come to bear on the choice of strategy providing the best fit for a given organisation. While implementation of any system designed to augment the computational support offered to metadata producers is far from trivial, we have demonstrated that the potential return on investment of effort can be considerable.

References

1. *Handling Geographic Information*; Department of the Environment, The report of the Committee of Enquiry chaired by Lord Chorley; Her Majesty's Stationary Office: London, UK, 1987.
2. Nanson, B.; Smith, N.; Davey, A. What is the British National Geospatial Database? In *7th Conference of the Association for Geographic Information*, Birmingham, UK, November 22, 1995; AGI: London, UK, 1995.
3. Davey, A.; Murray, K. Update on the National Geospatial Database - Collaboration between Organisations. In *8th Conference of the Association for Geographic Information at GIS96*, Birmingham, UK, September 24-26, 1996. AGI: London, UK, 1996.
4. Göbel, S.; Lutze, K. Development of meta databases for geospatial data in the WWW. In *ACM-GIS 1998, 6th International Symposium on Advances in Geographic Information Systems*, Washington, D.C., USA, November 6-7, 1998; ACM: New York, NY, USA; pp. 94-99.
5. Guptill, S.G. Metadata and data catalogues. In *Geographical Information Systems*; Longley, P., Goodchild, M.F., Maguire, D.J., Rhind, D.W., Eds.; Wiley: Chichester, UK, 1999; pp. 677-692.
6. Tsou, M.-H. An Operational Metadata Framework for Searching, Indexing, and Retrieving Distributed Geographic Information Services on the Internet. In *2nd International Conference on Geographic Information Science, GIScience 2002*; Egenhofer, M., Mark, D., Eds.; *Lecture Notes in Computer Science*, LNCS 2478, Springer-Verlag: Berlin, Germany, 2002; pp. 313-332.
7. Westbrooks, E.L. Distributing and Synchronizing Heterogenous Metadata for the Management of Geospatial Information Repositories for Access. In *Metadata in Practice: Building the Diverse Digital Library*; Hillmann, D., Westbrooks, E., Eds.; American Library Association: Chicago, IL, USA, 2004; pp. 139-157.
8. Boxall, J. Geolibraries: geographers, librarians and spatial collaboration. *Can. Geogr.* **2003**, *47*, 18-27.
9. Nogueras-Iso, J.; Zarazaga-Soria, F.J.; Lacasta, J.; Béjar, R.; Muro-Medrano, P.R. Metadata standard interoperability: application in the geographic information domain. *Comput. Environ. Urban* **2004**, *28*, 611-634.
10. Xie, R.; Shibasaki, R. Imagery Metadata Development Based on ISO/TC 211 Standards. *Dat. S. J.* **2007**, *6*, 28-45.
11. Hill, L.L.; Janée, G.; Dolin, R.; Frew, J.; Larsgaard, M. Collection Metadata Solutions for Digital Library Applications. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 1169-1181.
12. Tulloch, D.L.; Robinson, M. A progress report on a U.S. National Survey of Geospatial Framework Data. *J. Gov. Inform.* **2000**, *27*, 285-298.

13. Beaujardière, J.d.L.; Mitchell, H.; Raskin, R.; Rao, A. The NASA Digital Earth Testbed. In *ACM-GIS 2000, 8th ACM Symposium on Advances in Geographic Information Systems*, Washington, D.C., USA, November 10-11, 2000; ACM: New York, NY, USA; pp. 47-53.
14. Crompvoets, J.; Bregt, A.; Rajabifard, A.; Williamson, I. Assessing the worldwide developments of national spatial data clearinghouses. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 665-689.
15. Maguire, D.J.; Longley, P.A. The emergence of geoportals and their role in spatial data infrastructures. *Comput. Environ. Urban* **2005**, *29*, 3-14.
16. Nogueras-Iso, J.; Zarazaga-Soria, F.J.; Béjar, R.; Álvarez, P.J. OGC Catalog Services: a key element for the development of Spatial Data Infrastructures. *Comput. Geosci.* **2005**, *31*, 199-209.
17. Lutz, M.; Klien, E. Ontology-based retrieval of geographic information. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 233-260.
18. Buttenfield, B.P. Transmitting vector geospatial data across the Internet. In *Second International Conference on Geographic Information Science 2002*; Boulder, CO, USA, September 25-28, 2002; Egenhofer, M., Mark, D., Eds.; *Lecture Notes in Computer Science* LNCS 2478, Springer-Verlag: Berlin, Germany, 2002; pp 51-64.
19. Turner, A. *Introduction to Neogeography;* 2006. http://www.oreilly.com/catalog/neogeography/ (accessed 18 August 2009).
20. Yang, B.; Purves, R.; Weibel, R. Efficient transmission of vector data over the Internet. *Int. J. Geogr. Inf. Sci.* **2007**, *21*, 215-237.
21. Boucelma, O.; Colonna, F.-M. Mediation for Online Geoservices. In *Web and Wireless Geographical Information Systems: 4th International Workshop, W2GIS 2004*, Goyang, Korea, November 26-27, 2004; Claramunt, C., Kwon, Y.-J., Boujou, A., Eds.; *Lecture Notes in Computer Science* LNCS 3428, Springer-Verlag: Berlin, Germany, 2004; pp 81-93.
22. Agarwal, P. Ontological considerations in GIScience. *Int. J. Geogr. Inf. Sci.* **2005**, *19*, 501-536.
23. Peng, Z.-R. A proposed framework for feature-level geospatial data sharing: a case study for transportation network data. *Int. J. Geogr. Inf. Sci.* **2005**, *19*, 459-481.
24. Nebert, D.D. *Developing Spatial Data Infrastructures: The SDI Cookbook*; 2004, GSDI. http://www.gsdi.org/gsdicookbookindex.asp (accessed 18 August 2009).
25. Dunfey, R.I.; Gittings, B.M.; Batcheller. J.K. Towards an Open Architecture Vector GIS. *Comput. Geosci.* **2006**, *32*, 1720-1732.
26. Aloisio, G.; Milillo, G.; Williams, R. D. An XML architecture for high-performance web-based analysis of remote-sensing archives. *Future Gener. Computer Sy.* **1999**, *16*, 91-100.
27. Fonseca, F.T.; Egenhofer, M.J. Ontology-driven Geographic Information Systems. In *ACM-GIS 1999, 7th ACM International Symposium on Advances in Geographic Information Systems*, Kansas City, MO, USA, November 2-6, 1999; ACM: New York, NY, USA; pp. 14-19.
28. Kiehle, C. Business logic for geoprocessing of distributed geodata. *Comput. Geosci.* **2006**, *32*, 1746-1757.
29. Healey, R.G.; Delve, J. Integrating GIS and data warehousing in a Web environment: A case study of the US 1880 Census. *Int. J. Geogr. Inf. Sci.* **2007**, *21*, 603-624.
30. *OGC 2003 OpenGIS Reference Model (OGC 03-040).* http://portal.opengeospatial.org/files/?artifact_id=3836 (accessed 18 August 2009).

31. Higgins, C.; Medyckyj-Scott, D.; Reid, J. A Community Specific SDI - the Case of UK Academia. In *Geodaten- und Geodienste-Infrastrukturen - von der Forschung zur praktischen Anwendung*, Münster, Germany, June 26-27, 2003; University of Münster: Münster, Germany, 2003; pp. 77-89.
32. Balsoy, O.; Jin, J.; Aydin, G.; Pierce, M.; Fox, G. Automating metadata Web service deployment for problem solving environments. *Future Gener. Computer Sy.* **2005**, *21*, 910-919.
33. Greenberg, J.; Spurgin, K.; Crystal, A. Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions. *IJMSO* **2005**, *1*, 3-20.
34. Mathys, T. The Go-Geo! Portal Metadata Initiatives. In *12th Annual Conference of the GIS Research UK (GISRUK)*, Norwich, UK, April 28-30, 2004; University of East Anglia: Norwich, UK, 2004; 148-154.
35. Batcheller, J.K. Automating geospatial metadata generation—An integrated data management and documentation approach. *Comput. Geosci.* **2008**, *34*, 387-398.
36. Greenberg, J. Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. *JIC* **2004**, *6*, 59-82.
37. AGI 2004 UK GEMINI Standard Version 1.0 - A Geo-spatial Metadata Interoperability Initiative. http://www.govtalk.gov.uk/schemasstandards/metadata_document.asp?docnum=903 (accessed 18 August 2009).
38. Kokla, M.; Kavouras, M. Fusion of top-level and geographical domain ontologies based on context formation and complementarity. *Int. J. Geogr. Inf. Sci.* **2001**, *15*, 679-687.
39. Hunter, G.J. Spatial data quality revisited. In *Proceedings of GeoInfo 2001*, Rio de Janeiro, Brazil, October 4-5, 2001; pp. 1–7.
40. Devillers, R.; Gervais, M.; Bédard, Y.; Jeansouin, R. Spatial Data Quality: from Metadata to Quality Indicators and Contextual End-User Manual. In *OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management*, Istanbul, Turkey, March 21–22, 2002; pp. 45-55.
41. *ISO 2005 Geographic Information – Metadata*; BS EN ISO 19115, BSi British Standards: Bristol, UK.
42. *ISO 2005 Geographic Information - Rules for Application Schema*; BS EN ISO 19109, BSi British Standards: Bristol, UK.
43. Comber, A.J.; Fisher, P.F.; Wadsworth, R.A. User-focused metadata for spatial data, geographical information and data quality assessments. In *10th AGILE International Conference on Geographic Information* Science, Aalborg, Denmark, May 8-11, 2007.