Bridging the Gap between Traditional Metadata and the Requirements of an Academic SDI for Interdisciplinary Research

Claire Ellul[1], Daniel Winer[1], John Mooney[2] and Jo Foord[2]

[1]University College London
[2]London Metropolitan University

c.ellul@ucl.ac.uk; daniel.winer.10@ucl.ac.uk; j.mooney@londonmet.ac.uk; j.foord@londonmet.ac.uk

Abstract

Metadata has long been understood as a fundamental component of any Spatial Data Infrastructure, providing information relating to discovery, evaluation and use of datasets and describing their quality. Having good metadata about a dataset is fundamental to using it correctly and to understanding the implications of issues such as missing data or incorrect attribution on the results obtained for any analysis carried out.

Traditionally, spatial data was created by expert users (e.g. national mapping agencies), who created metadata for the data. Increasingly, however, data used in spatial analysis comes from multiple sources and could be captured or used by non-expert users – for example academic researchers - many of whom are from non-GIS disciplinary backgrounds, not familiar with metadata and perhaps working in geographically dispersed teams. This paper examines the applicability of metadata in this academic context, using a multi-national coastal/environmental project as a case study. The work to date highlights a number of suggestions for good practice, issues and research questions relevant to Academic SDI, particularly given the increased levels of research data sharing and reuse required by UK and EU funders.

KEYWORDS: metadata, Spatial Data Infrastructures, GIS, inter-disciplinary, cross-disciplinary, Academic SDI

1. Introduction

Until the emergence of the geographical information technologies that are part of Web Mapping 2.0 (Goodchild, 2007, Haklay *et al.,* 2008, Elwood, 2009) geographical information was provided top-down by bodies such as National Mapping Agencies (NMA) (Goodchild in Schuurman, 2009). Advances in positioning, web mapping, cell/mobile communications, Web 2.0 and Volunteered Geographic Information (VGI) (Goodchild, 2007b) have led to increasing availability of data from multiple sources (Budhathoki *et al.,* 2008), with much of this spatial data available free of charge (Coleman *et al.,* 2009).

In the context of academic research in the United Kingdom (UK), a number of measures have responded to, and reflect, this greater availability of data. At European level, Seventh Framework Programme (FP7) funding requires funded projects to provide a data management plan (FP7, 2011) and the European Union's INSPIRE (Infrastructure for Spatial Information in Europe) (INSPIRE, 2011a) directive may impact academia. Initiatives to encourage greater sharing of research data are being established – e.g. the Engineering and Physical Sciences Research Council's *Policy Framework on Research Data* (EPSRC, 2011) and the Economic and Social Research Council's *Research Data Policy* (ESRC, 2010). The setting up of an Academic Spatial Data Infrastructure (SDI) is one of the aims of the Joint Information Systems Committee's Geospatial Working Group (JISC, which was set up to facilitate information and infrastructure sharing across the UK's universities) (JISC, 2011). Initiatives such as GoGeo[1] and ShareGeo[2] allow academic users to share geospatial data online.

This increase in available spatial data is coupled with a reduction in Geographic Information System (GIS) expertise of the end user of such data. Previously, users were GIS experts with advanced training in spatial data understanding and management and quality issues. However, the British Library recently predicted an increasing emphasis on cross-disciplinary research (British Library, 2010). Initiatives such as research projects funded by the JISC Geospatial Programme (JISC, 2011b) recognize the importance of spatial data analysis and GIS to other disciplines. The availability of free GIS software (e.g. Google Maps[3], Google Earth Builder[4], ArcGIS Explorer[5], ESRI's Community Analyst Tools[6], Quantum GIS[7]) encourages non-specialist

[1] http://www.gogeo.ac.uk/gogeo/
[2] http://www.sharegeo.ac.uk/
[3] http://maps.google.com
[4] http://earth.google.com/builder
[5] http://www.esri.com/software/arcgis/explorer/index.html
[6] http://www.esri.com/software/arcgis/community-analyst/index.html

users to make use of GIS tools and data. This is particularly the case given the power of GIS as a tool for the integration of data from diverse sources and disciplines.

Given both the increase in data and the reduction in expertise of the users, having information to allow end-users to understand and integrate the heterogeneous data they are using, and identify any potential issues, omissions, data capture methods and previous analysis carried out, becomes more important (Deng and Di, 2009, Haklay and Weber, 2008). Traditionally, metadata ('data describing the data') has been used (Sboui *et al.,* 2009) and amongst the GIS profession the quality description provided by metadata is acknowledged as important to understand potential errors and issues. Good metadata increases trust (Craglia *et al.,* 2008) and could be important to help increase the credibility of a dataset, mentioned by Coleman *et al.* (2009) as important particularly for VGI. However, metadata is complex to create (Poore and Woolf 2010, Manso-Callejo *et al.,* 2010) and "many view its generation as monotonous and time-consuming" (Batcheller, 2008), standards are producer-centric (Goodchild, 2007, Devillers *et al.,* 2005) and where metadata exists its quality may be variable (Rajabifard *et al.,* 2009). Indeed, many systems currently rely on "caveat emptor" (Goodchild, 2007).

This paper describes a review of metadata creation and use in a multi-national, interdisciplinary research project where the data quality description it provides is fundamental to the success of the project. The review examines whether traditional metadata, as a descriptor of data quality, is relevant to and usable in an Academic SDI and if there are any considerations that could overcome some of the issues commonly associated with its use.

The remainder of the paper is structured as follows – first a review of data quality issues and metadata is given. This is followed by an overview of the case study (the SECOA project). The results of an evaluation of SECOA's use of metadata are then presented, along with consideration as to whether metadata is relevant and usable for academic research. The paper concludes by presenting some ideas and concepts for further work to more tightly integrate metadata into the academic data management workflow.

2. Data Quality and Metadata

Concerns about accuracy and uncertainty of geographical datasets have been articulated for some time (Goodchild, 2002*)*. The level of vagueness (zone boundaries are possibly guesses), uncertainty (both positional and attribute) and ambiguity (e.g. where objects are assigned different labels by different groups or disciplines) (Longley *et al.,* 2011) all contribute to the quality of a dataset. Borrough (1994) lists potential sources of error in data including the age of the data, areal coverage, map scale, density of observations, relevance, format and accessibility. Van Oort (2006) identifies

[7] http://www.qgis.org/

a number of groupings of geospatial data quality information: lineage (the history of the dataset, how it was collected, and how it evolved); positional accuracy (how well the coordinate value of an object in the database relates to reality on the ground); attribute accuracy (how correct attribute values are); logical consistency (does the dataset conform to rules such as 'no houses in the middle of a lake' and general topological correctness and other relationships that are encoded in the database); completeness (is there any missing data or any data included that should not be there); semantic accuracy (how should objects in the dataset be interpreted); usage (how the data should be used appropriately); temporal quality (if the real world changes, does the dataset change too?).

Within GIS, and in particular within an SDI it is the metadata that provides a formal description of the data quality (Kim, 1999), allows for data reuse (Craglia *et al.,* 2008) and avoids data duplication. To enable interchange and understanding by computer-based systems, metadata is often stored in a very structured, standardized format (e.g. the United States Federal Geographic Data Committee[8] or the International Standards Organization's 19115:2003 Geographic Information Metadata Standard[9]). A study by Moellering (2005) identified 22 standards still in wide use. Table 1 below lists core elements of metadata for the European Union's INSPIRE Spatial Data Infrastructure (INSPIRE, 2011b). As can be seen the information stored in standards-based metadata directly corresponds to the list of quality elements identified above, with additional information to facilitate searching for the dataset and sourcing it once its quality has been evaluated.

Metadata Element	**Metadata Element**	**Metadata Element**
Title	Data format	Extent
Alternative title	Responsible organization	Vertical extent
Dataset language	Frequency of update	Spatial reference system
Abstract	Limitations on public access	Spatial resolution
Topic category	Use constraints	Resource locator
Keyword	Additional information Source	West bounding longitude
Temporal extent	Metadata date	East bounding longitude
Dataset reference date	Metadata language	North bounding latitude
Lineage	Metadata point of contact	South bounding latitude
Originating controlled vocabulary	Unique resource identifier	Coupled resource

Table 1. INSPIRE Metadata Elements (adapted from Walker, 2009)

Traditionally, metadata is created by a dedicated team of professionals (Mathes, 2004 in Kalantari *et al.,* 2010, Budhathoki *et al.,* 2008) and metadata standards are producer centric (Goodchild, 2007, Devillers *et al.,* 2005, Craglia *et al.,* 2008). They focus on

[8] http://www.fgdc.gov/metadata
[9] http://portal.opengeospatial.org/files/?artifact_id=6495

information that data producers assume will be relevant to users and it is difficult for end-users to be involved at any point (Budhathoki *et al.,* 2008). These geospatial specialists understand the importance of producing and maintaining metadata and the underlying requirement to provide quality information with a dataset to ensure that it is used correctly for any subsequent analysis (Sboui *et al.* 2009). However, even for specialists the complexity of creating and maintaining such metadata is considered significant (Poore and Woolf, 2010, Manso-Callejo *et al.,* 2010, Batcheller, 2008, Craglia *et al.,* 2008). Metadata production is seen as tedious and left to the end of a project, which results in metadata that is barely useful and often contains errors (West and Hess, 2002).

Two approaches can be identified to automatic metadata production. First, it may be possible to automate data quality assessment and hence generate metadata from the results. This has been attempted by comparing the data with 'better/higher' quality datasets (Koukoletsos *et al.,* 2011) and through modeling (deBruin, 2008, Agumya and Hunter, 2002) and through examining the different values of nominal, ordinal, ratio and interval data (Van Oort, 2006, Servigne *et al.,* 2006). Secondly, direct automated metadata creation has also been attempted. Potential approaches here include harvesting existing metadata (Batcheller, 2008), automated tagging (Kalantari *et al.,* 2010), title and location information extraction (Olfat *et al.,* 2010), format, number and types of geometry, resolution, bounding box, use constraints (Manso-Callejo *et al.,* 2009). However, elements of metadata – in particular descriptions such as abstracts - creation cannot ever be eliminated from the process (Batcheller, 2008).

In addition, end-users may require further non-standard information. For example, they may wish to express their own measures of fitness-for-purpose (Craglia *et al.,* 2008), to add information providing a simple description of data quality or details of the impact that the dataset could have on the outcome of any analysis they wish to perform (Goodchild, 2007) or to describe data in terms aimed at non-expert users (Timkpf *et al.,* 1996, Frank, 1998 and Harvey, 1998 in Devillers *et al.,*2005). Poore and Wolfe (2010) note that issues relating to semantics and ontologies are not handled by current standards. Devillers *et al.* (2005) mention that the reputation of the data producer is important. Legal requirements are suggested as being relevant by Gervais (2004 in Devillers *et al.,* 2005) and Aalders and Morrison (1998 in Devillers *et al.,* 2005) propose including information about where a dataset has been used.

3. The SECOA Project

SECOA (Solutions for Environmental Contrasts in Coastal Areas) is a research project involving eight different universities and institutions around the world (in the United Kingdom, Italy, Portugal, Israel, India, Vietnam, Sweden and Belgium). It has been set up to examine the effects of human mobility on urban settlement growth and in fragile environments – in particular the potential impact of sea level rise (SECOA, 2011a; 2011b). SECOA is investigating and comparing eight metropolitan areas of international/global importance and an additional eight metropolitan areas of

regional/national importance in these European and Asian countries. Given the wide range of issues to be addressed by the project, the SECOA team recognized the importance of data and data management from the outset. Metadata forms a core component of the data management task and specific time for metadata capture was allocated in the project schedule.

SECOA's metadata end-users can broadly be divided into three groups: producers (creating metadata and datasets for others), users (making use of metadata and datasets for cross-location comparison and model building) and "produsers" (given the small teams, a number of people fell into both roles). The teams are very interdisciplinary and include researchers having expertise in the Creative Industries, Fluvial and Flood Geomorphology, Tourism Studies, Urban Planning, European Integration and Globalization among others.

Although standards-based metadata (in particular INSPIRE) was considered at the outset of the project, its complexity resulted in the creation of a shorter version of metadata ("stripped down", Longley *et al.,* 2011) to describe the datasets and be manageable in terms of creation time and understanding by the end users. The required metadata fields were identified through a questionnaire issued to the end users themselves (see Figure 1 below). Importantly, flexibility was included – users could upload documents to provide more detailed data quality information, and additional elements of metadata can be added as the project progresses, building towards the INSPIRE standard (see Ellul *et al.,* 2009 for details of how this is achieved). To assist the metadata creation task detailed guidance was produced in the form of user guides, decision flow diagrams and example metadata records. To address the issue of the diverse backgrounds of the team, regular presentations to familiarize users with metadata and data management are given at the six-monthly project meetings. At all times, the emphasis is on the use of metadata as a means to allow users to correctly and scientifically use, integrate and compare datasets from multiple sources and for multiple locations.

Throughout the first eighteen months of project activity, usage of the system has been tracked – users' requests for metadata have been logged, along with the number of metadata records and associated data files uploaded – to provide a quantitative insight into the system. Additionally, a qualitative review of metadata captured has been carried out to assess the usage and perceptions of the metadata system from the perspective of content.

4. The SECOA Metadata System – Results

Figure 1 shows the resulting web-based metadata system, with the elements highlighted by producers, produsers and users as important.

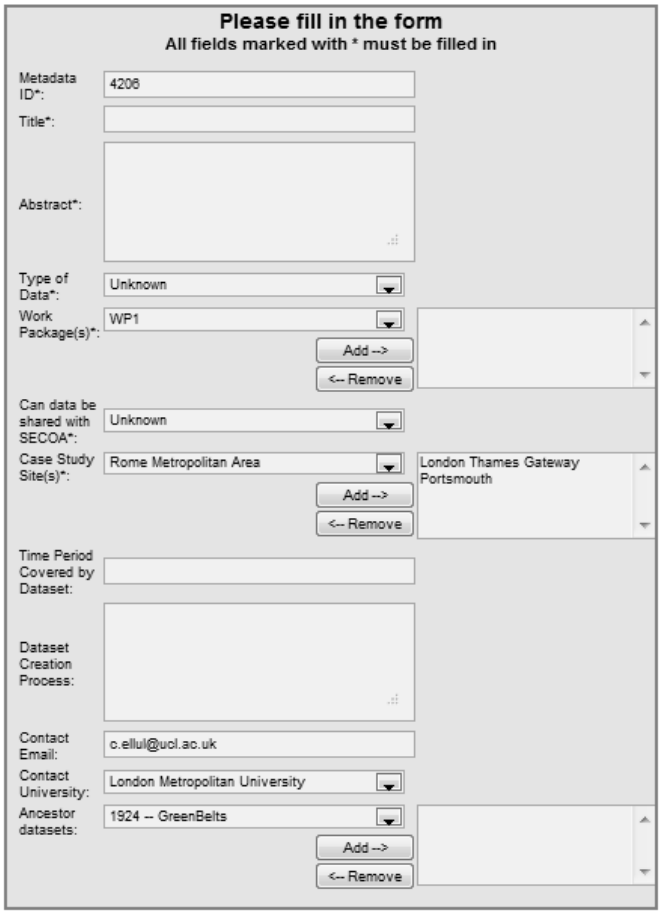

Figure 1. SECOA Metadata Capture Form

In the above Figure, the following elements of metadata have been included: a short *Title* (around 5 words) that describes the dataset; an *Abstract* to give a short description of the dataset; the *Type of Data* – such as spreadsheet, spatial data, PDF; the *Time Period(s) covered by the data* – of particular importance given the time-based change analysis in SECOA; *How the dataset was created* – details to allow the user of the dataset to understand how particular numbers or results were derived' the relevant SECOA *Work Packages*; whether *Data can be shared with SECOA.* Items such as *Contact E-mail*, relevant *Case Study* and *Contact University* are captured automatically from the user's login. Additionally, the system provides the ability for

users to link to *'ancestral datasets'* if a dataset is derived from another, to upload additional files describing the data and to upload the data file itself where it can be freely shared.

4.1 Quantitative Evaluation of the SECOA System

Figure 2 below shows the number of metadata records created by each of the partners (anonymized except for London Metropolitan University, LMU, the creators of the metadata system).

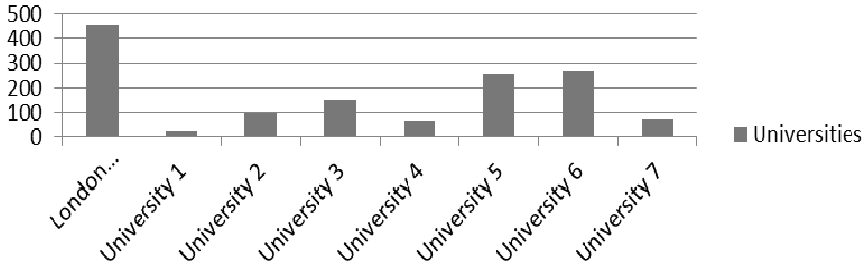

Figure 2. Number of Metadata Records Uploaded by Each Partner University

A total of approximately 1800 records have been created to date (October 2011). However, as can be seen, there has been a mixed response to the system with University 1 having submitted little metadata despite repeated encouragement, but others (5, 6 and LMU) performing well. Additionally, a total of 545 files (containing data or additional metadata information) have been uploaded.

Figure 3 below examines usage of the system for metadata viewing, again by anonymized university, with LMU excluded from the list. There have been approximately 2800 individual views of metadata records by non-LMU staff since system launch, but again there is great disparity between the teams with the universities showing a good record for metadata population also showing a good record for general use of the system. Detailed tracking results also show that there are relatively few users accessing the system in a significant way in each location - 13 core users (outside LMU) exist, who have viewed over 100 metadata records each since the system was launched.

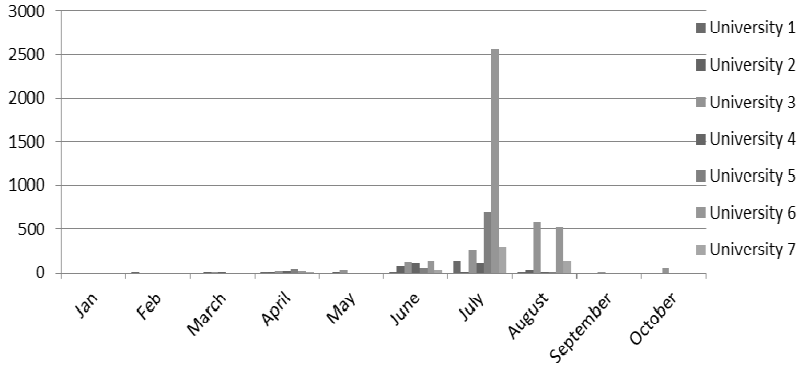

Figure 3. Number of Metadata Records Viewed by Partner Universities, by Month

The importance of a deadline in encouraging metadata submission cannot be underestimated - an additional 715 records were created in July 2011 in anticipation of the first metadata deadline. This is reflected in the heavier system usage in July in Figure 3 above.

4.2 Reviewing the Quality of SECOA Metadata

The disparity in the number of metadata records captured by the teams highlighted an issue with inconsistent metadata with some teams missing records although in theory all teams were required to contribute the same analysis results and associated metadata records to the project to allow comparability across the countries. A review of metadata content also highlighted the great variety of detail present in the metadata. For example time-periods covered by various datasets included "1915 to present - variable depending on the location", "Collation of data as in Jan 2009" and "Details attached - depends on data type". Different descriptions and levels of detail were provided for data for a requested Driving forces, Pressures, States, Impacts, Responses (DPSIR) report (records below are anonymized):

- *University 1 created one metadata record with the abstract details: "DPSIR framework analysis for ecosystem of City A and City B"*
- *University 2 created one metadata record with abstract details: "Assessment of natural resources use for sustainable development (DPSIR analysis). The coastal wetlands in the municipalities of City A (peri-urban area) and City B (peri-urban area)"*
- *University 3 created eight metadata, with abstract details: "Report on the assessment of sustainable use of natural resources in the City A study sites: District and District B. The DPSIR framework is used to assess the sustainability of intertidal habitats in six statutory conservation areas. An index of sustainability is developed based on eight selected indicators. Results are very dependent on the*

indicators used and their relative weight. Therefore the index is used here only to rank the six areas based on the relative level of pressure they currently "

Provision of more detailed guidance for metadata capture is on-going. First, a decision tree is sketched out to allow users to determine whether a metadata record is required to be captured or not (Figure 4).

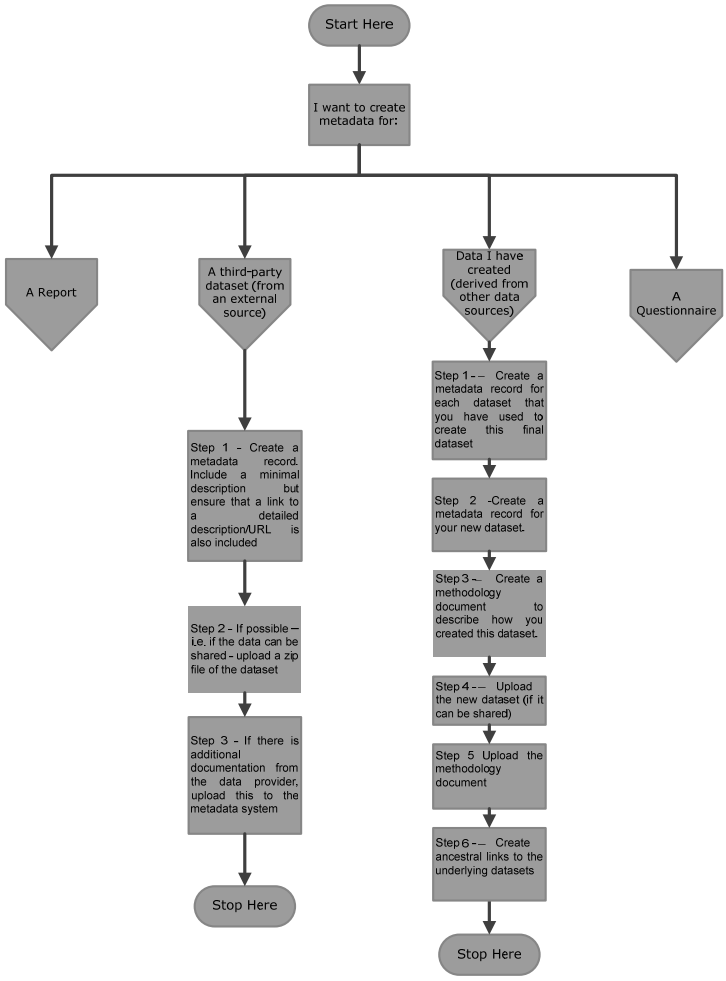

Figure 4. Decision Tree Diagram Guidance for Metadata Capture

Secondly, a series of best-practice examples have created in the metadata system by the LMU team.

A second issue to be addressed is how best to assess the quality of the metadata produced in an automated fashion. Although the manual review described above yields relevant results, this is not scalable to hundreds or even thousands of user-generated metadata records. A quality assessment measure was therefore applied to the metadata, using the following criteria:

- The total amount of text provided in the abstract
- The total amount of text provided for the description of the dataset creation process
- The links between each metadata record and parent records.

Figure 5 below shows early stage results of this type of analysis, with 15 being a maximum quality score for a metadata record. The analysis highlighted in particular the lack of 'links' to parent datasets and the lack of text in some metadata entries. Individual reports will be circulated to all participants to encourage them to improve their scores.

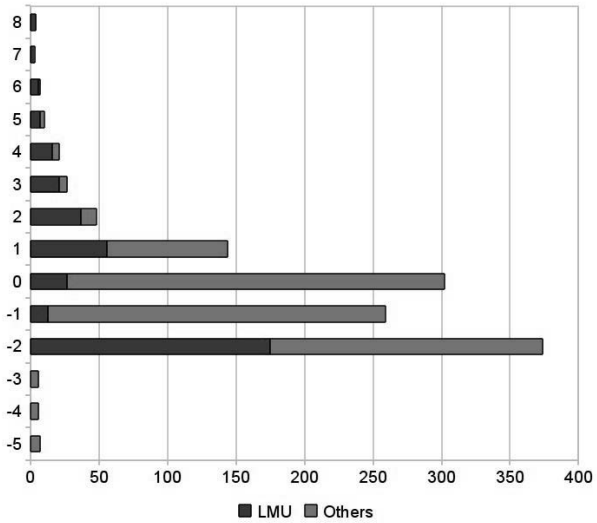

Figure 5. Automated Metadata Quality Analysis

4.3 Qualitative Evaluation of the Usage of the SECOA System

A second short questionnaire relating to usage of the metadata system yielded a total of 10 responses from users (5 out of 8 countries responded). Users were asked what they were using the system for, and whether they managed to locate the data they needed for their analysis work. Responses are given in Table 2 and Table 3.

Why do you use the Metadata System and Forum?	# Responses
To Upload metadata and datasets	8
To discuss issues on the forum	6
To search the metadata	6

Table 2. Metadata and Forum Usage

Do you find all the metadata/data that you need in the system?	# Responses
Yes - I find all the metadata/data I needed	3
Sometimes some metadata/data is missing	4
No - I cannot find what I need	0
I am not using the SEARCH option in the metadata tool	3

Table 3. Metadata Completeness

Two of the respondents, both members of the team currently conducting comparative studies, identified specific areas of missing metadata (and hence data that they required for analysis). Other issues included occasions where data did not meet the requested format (e.g. a PDF was supplied instead of a spreadsheet).

5. Is Metadata Usable and Useful within an Academic Research SDI?

Overall, the total of 1800 metadata records and 545 datasets uploaded and shared by the SECOA team point to a general level of success of the metadata tools. Having real, project-related, deadlines and having the data repository (and hence metadata) as an external deliverable with specific person-months allocated to it in the project schedule was fundamental to reaching this level of metadata as this gave the task higher impact. The majority of the work was carried out by a core team of 10 users, who have created on average 150 records each and quantitative assessment, by means of usage logging highlighted that within each team there are usually one or two 'metadata champions' who perform the majority of the entries and searches on behalf of the team.

The introduction of the "stripped down" metadata capture requirements and the automation of metadata capture for a number of elements was particularly successful, as was making the users aware that they would not be required to populate complex, complete standards-based metadata. Given the low level of individual queries to the development team (perhaps 5-10 across the first year) it would appear that the web-based system provided (along with the associated instructions) was deemed usable.

Members of the project team have become more familiar with metadata as the project has progressed. Feedback from the end users of the metadata – those team

members using the captured metadata – is also positive overall. Users were generally able to locate the datasets they required using the system's search tool, and the geographically dispersed project teams means that metadata was a first port of call for the teams searches for data, rather than an e-mail or phone call to the relevant team member. This use of metadata was also relevant within teams - anecdotally people were able to use metadata to answer questions about the datasets where details may have been forgotten due to elapsed time. Where clarification has been necessary, it has been possible to ask people to go back and add to or improve their metadata.

Comparing SECOA to traditional SDI, it can be realized that SECOA uses not only metadata on 'official' data but also requires metadata for the aggregated/analyzed data produced for comparative analysis. The metadata reflects the different methods used to produce the aggregated data, allowing comparison between the results from different teams. It is noteworthy that the results were often not spatial in nature, but consisted of summary reports or spreadsheets of aggregated numbers. Thus the SECOA SDI, and perhaps research SDI in general, needs to be able to handle both spatial and non-spatial data.

Despite the successes a number of issues have emerged which can be said to reflect those identified above (*Data Quality and Metadata*). Users have noted that some datasets and metadata are missing (i.e. have not been created/uploaded as required by various country teams) and our review highlighted inconsistent metadata creation and great inconsistencies in the resulting quality of the metadata. The SECOA team also exhibited the behaviour often described in association with metadata, where metadata was ignored in favour of more pressing data capture and analysis deadlines, unless specific metadata deadlines were set, and it remains to be seen whether participants will be willing to go through additional iterations to improve the quality of the metadata created.

Importantly, SECOA illustrates that metadata is relevant to facilitate data sharing and data quality description and ultimately ensure better science. Ideally, metadata and the data repository would be an external deliverable, and it is suggested that metadata deadlines are set on a frequent basis and accompanied by metadata review exercises. The issues with the quality of the metadata highlight the need for multiple iterations of metadata creation and maintenance to be scheduled and costed, and the need for detailed guidance and examples to be pre-created.

The time required to create detailed, more consistent, high quality metadata, perhaps including additional non-standard elements (see *Data Quality and Metadata* above), should not be underestimated. Even if, as was the case with SECOA users contribute fairly extensive metadata they are predominantly not GIS experts. Do they have the expertise in spatial data sufficient to do so with sufficient understanding of the limitations of their datasets? Therefore, perhaps the most fundamental question to address is 'how can we automate metadata capture and data quality assessment and documentation?' If data has been manipulated or analyzed in a GIS, the metadata could list the software package and version, and also the exact operations that were

performed, in order, information which would not only be useful for the project but would contribute to the repeatability of the research downstream. However, even given this level of automatic data quality/metadata creation, fully automated metadata is as yet unreachable.

An interim alternative could be proposed that incorproates metadata directly into a user's workflow – in other words, datasets cannot be accessed (e.g. in the GIS) without the user being made aware of corresponding metadata and hence any data quality issues, and cannot be shared without appropriate metadata being created (this contrasts with current systems, where metadata is held separately). Storing metadata with the data in an integrated single environment such as a spatial database would greatly assist in enforcing such rules. It would also allow the system to automatically update the metadata when the underlying datasets change (by means of a 'trigger' event in the database) and could generate regular prompts to the user to ensure that the metadata was up to date. Logging of GIS operations could be done directly into the database, and metadata records would be automatically created for new datasets, reducing the need for guidance and the existence of a separate 'metadata creation' task. Text mining tools could be used to automatically detect abbreviations and flag them to the user if they are not already logged in the system. Voice recording and transcription services could be included to faciliate the population of mandatory elements that cannot be automated, such as *title* and *abstract*.

The above measures may go some way to overcoming the wider issue of the complexity (and relevance) of standards-based metadata and the general perception that it is 'boring', 'irrelevant' and 'difficult to create and use' (Pasca *et al.,* 2009). To further this process, consideration should be given once again to one of the main purposes of metadata – it is a representation of the quality of the data, and should flag up any issues relating to the dataset to potential end users, empowering them to source data, make a descision as to whether to use a dataset and if used how to interpret the results obtained. Familiarizing researchers with the importance of such data quality descriptions to their project could assist in this task. Understanding motivation (from altruism to social reward, as suggested by Coleman *et al*. 2009) is relevant, as are participative methods of user feedback (Craglia *et al.,* 2008).

From the metadata creation perspective, techniques could involve adding quality ratings and descriptions to be applied both to the datasets and to the metadata - "*I used this dataset for task XYZ*", "*I rank this dataset as 4/5*", "*I found these issues in this data*", "*The metadata failed to mention that there is an entire county missing in the data*." Further research into the applicability of the initial quality measures used above (Section 4.2) is also required – how can the quality of large numbers of metadata records be assessed on an ongoing basis? Online games could be created, with users competing in teams to describe spatial datasets and identify the most appropriate tags. More generally, the following questions 'how can we highlight the importance of understanding data quality?' and 'what would motivate people to voluntarily contribute metadata/quality information?' are relevant.

From the metadata user's perspective it is equally important to ensure that the resulting quality descriptions are relevant, and used in the correct context. Do users of metadata, increasingly not GIS experts, have the skills to interpret its meaning in terms of the underlying data quality and its impact on their analysis? 'How can people be encouraged to make use of metadata to obtain data quality information and correctly interpret the impact of data quality on their analysis and results?'

Automation has been discussed in the context of metadata creation, and it is possible that it may play a part here too, realizing one of the advantages of the structured approach to metadata storage. Given that it is created in a format to be machine-readable could such metadata be used to automatically assess the suitability of a dataset for a specific task, or perhaps issue warning flags or descriptions of 'suitable' datasets? For example, what is an appropriate point density for an inverse distance weighting interpolation with particular parameters? Does the proposed dataset have this appropriate point density? This concept extends the concept of metadata to processes and algorithms - a metadata record of an 'ideal' dataset could be created for each task, and then compared to that of the proposed dataset. Given the wider audience now using GIS (see *Introduction*) this would help to ensure that appropriate scientific output was produced and add an increased level of usability for novice users.

5. Conclusions and Further Work

The SECOA project could be said to reflect data creation and management requirements occurring across interdisciplinary, multi-national research and Table 4 highlights a number of similarities and differences between a 'traditional' SDI as exemplified by INSPIRE and an 'academic' SDI.

'Traditional SDI'	'Academic SDI'
Complex metadata standards	Stripped-down metadata standards, but may have additional non-standard extras such as 'ancestor links', 'work package' or ratings.
Designed to handle spatial data only	Needs to handle mixed data including spatial, reports, questionnaires
Producer centric, data provided to anyone who requests/licenses it.	Both producer and user centric, as well as produsers. Data shared within a project, although greater emphasis now emerging on longer data life-cycle.
Expert producers, expert users who understand the importance of metadata and the detailed level of metadata required	Non-expert producers and users, who are not familiar with metadata and may not have expertise in interpreting it and then applying this interpretation to their research
Deadlines for metadata production	Deadlines for metadata production only exist if set within the initial project scope
Multi-Lingual metadata	Generally a single language agreed for each project, although multi-lingual also possible.
Ongoing data updates and metadata maintenance	Data updates and metadata maintenance end with the individual project.
Domain expertise high – e.g. many data producers participate in the working groups that define the standards for the data and metadata in their area of expertise	Metadata and data domain expertise can be very low – academics are generally specialists in their own field, rather than in data management. Important to familiarize team members with metadata concepts early on. Metadata champions important.
Time is allocated to metadata production	Time is only allocated to metadata production if defined as part of the original project scope.
Quality of metadata generally good – producers of the metadata know their data well	Quality of metadata can be poor, and metadata can be missing. Difficult for non-metadata experts to understand how much detail to provide. Further methods required to automatically understand the quality of metadata.
Metadata held separately from data	Metadata held separately from data. Ideally creation of quality information and application of this information to subsequent analysis should be integrated into the workflow and potentially 'hidden' from the end users.
Metadata time consuming to produce.	Metadata time-consuming to produce, automation fundamental to resolving this issue.

Table 4. Traditional Versus Academic SDI

The SECOA project is currently two-years into a four-year timescale. As well as ongoing quantitative measurements such as those described above (Quantitative Evaluation of the SECOA System), producers, produsers and users of the metadata system will be

surveyed again to identify issues, successes and their overall level of understanding of metadata. Lessons learned from SECOA, such as the importance of familiarizing end users with metadata early on and the importance of including metadata as a deliverable, can be directly applied to further interdisciplinary research and a more integrated spatial database and metadata system is currently being developed for another project.

Metadata is an established means to convey the quality of a spatial dataset and allow the user to locate data, understand its suitability for a task, undertake the required analysis and release and share the results. On the one hand, traditional standards-based metadata provides a potential opportunity to semi-automatically assess the suitability of a dataset for a specific task. Conversely, the complexity of such metadata (and the omission of more end-user-focused concepts such as a quality rating from the standards) discourages its creation and maintenance. Many challenges remain, both for SECOA and the wider world of Academic SDI in an increasingly inter-disciplinary and geographically dispersed research context, not the least of which is identifying a suitable descriptor or set of descriptors for data quality that are both easy to create (at least semi-automatically) and relevant to end-users. If the process can be simplified for both metadata generation and search, inexperienced users will be more likely to use such systems and in doing so there should be an increase in the cooperation between research and a reduction in the cost of unnecessary and repeated research (EPSRC, 2011).

The current trends in GIS – increasing amounts of freely available data and web-based and desktop processes and software, along with an increasing user base of non-specialists, have major implications for geospatial scientists. Ensuring that non-experts make informed, correct and scientific choices of data and relevant operations has implications for the quality of the resulting output and the reputation of the discipline as a whole. Education forms a key part of this, and the developers of training material for non-specialists should ensure that issues relating to data quality are included. In an ideal world, such metadata would be seamless and hidden. However, the data quality and the implications of quality on analysis would be displayed more prominently than in current tools.

Acknowledgements

The authors would like to acknowledge the contributions made by the SECOA team to this research, part of which has been funded by the European Union's Framework 7 programme: Project n°: 244251 FP7-ENV.2009.2.1.5.1. We would also like to thank the participants at the recent European Science Foundation exploratory workshop on Laser Scanning Spatial Data Infrastructures (Heidelberg, 8-10 September 2011), for their insight into the broader reach of modern SDI.

References

Agumya, A. and Hunter, G., (2002), Responding to the consequences of uncertainty in geographical data, International Journal of Geographic Information Science 16(5): 405-417.

Batcheller, J., (2008), Automating geospatial metadata generation – An integrated data management and documentation approach, Computers & Geosciences, 34: 287-398.

Borrough, P., (1994), Principles of Geographical Information Systems for Land Resources Assessment, Oxford Science Publications, Clarendon Press, Oxford.

British Library (2010), 2020 Vision Project -Trends in Universities, Research and Higher Education - Internal discussion paper [online]. Available from: http://www.bl.uk/aboutus/stratpolprog/2020vision/trendsinuniresearch3.pdf 2010 [Accessed 21st November 2011].

Budhathoki, N.R., Bruce, B., (Chip), and Nedovic-Budic, Z., (2008), Reconceptualizing the role of the user of spatial data infrastructures, GeoJournal: An International Journal on Geography, 72(3-4): 149-160.

Coleman, D.J., Georgiadou, Y., and Labonte, J., (2009), Volunteered Geographic Information: The Nature and Motivation of Produsers ,International Journal of Spatial Data Infrastructures Research 4: 332-358.

Craglia, M., Goodchild, M. F., Annoni, A., Camara, G., Gould, M., Kuhn, W., Mark, Masser I., Maguire, D., Liang, S., and Parsons, E., (2008), Next Generation Digital Earth. A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science, International Journal of Spatial Data Infrastructures Research 3:146-167.

De Bruin, S., (2008), Modelling Positional Uncertainty of Line Features for Stochastic Deviations from Straight Line Segments, Transactions in GIS, 12(2): 165-177.

Deng, M., and Di, L., (2009), Building an Online Learning and Research Environment to Enhance Use of Geospatial Data, International Journal of Spatial Data Infrastructures Research.

Devillers, R., Bedard, Y., and Jeansoulin, R., (2005), Multidimensional Management of Geospatial Data Quality Information for its Dynamic Use Within GIS, Photogrammetric Engineering & Remote Sensing, 71(2): 205-215.

Ellul, C., Haklay, M., Francis, L., and Rahemtulla, H., (2009), A Mechanism to Create Community Maps for Non-Technical Users", IEEE Computer Society, GEOWS, International Conference on Advanced Geographic Information Systems & Web Services, pp 129-134.

Elwood, S., (2009), Geographic Information Science: New geovisualization technologies – emerging questions and linkages with GIScience Research, Progress in Human Geography 33(2): 256-263.

EPSRC (2011), EPSRC Policy Framework on Research Data. [online] Available from: http://www.epsrc.ac.uk/about/standards/researchdata/Pages/default.aspx [Accessed 21st November 2011].

ESRC (2010), Research Data Policy. [online] Available from: http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf [Accessed 21st November 2011].

FGDC (2010), Geospatial Metadata Tools – Federal Geographic Data Committee. [online] Available from: http://www.fgdc.gov/metadata/geospatial-metadata-tools [Accessed 24th March 2010].

FP7 (2011), Research Data Management – European Commission – FP7 [online] Available from: http://www.admin.ox.ac.uk/rdm/managedata/funderpolicy/ec/ [Accessed 21st November 2011].

Goodchild, M. F., (2002), "Introduction to Part I: Theoretical models for uncertain GIS", in W. Shi, P.F. Fisher, and M.F. Goodchild, editors, *Spatial Data Quality*. New York: Taylor and Francis, pp. 1–4. [367].

Goodchild, M. F. (2007), "Beyond Metadata: Towards User-Centric Description of Data Quality", Paper read at Proceedings, Spatial Data Quality 2007 International Symposium on Spatial Data Quality, June 13-15, at Enschede, Netherlands.

Goodchild, M. F., (2007b). Citizens as sensors: the world of volunteered geography. *GeoJournal* 69: 211–21.

Haklay, M., Singleton, A.D., and Parker, C., (2008), "Web mapping 2.0: the Neogeography of the Geospatial Internet", *Geography Compass*, (3).

Haklay, M., and Weber, P., (2008), "OpenStreetMap – User Generated Street Map", *IEEE Pervasive Computing.* October-December 2008, pp. 12-18.

INSPIRE (2011a), About Inspire [online] Available from: http://inspire.jrc.ec.europa.eu/index.cfm/pageid/48 [Accessed 23rd March 2011].

INSPIRE (2011b). INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119 [online] Available from: http://inspire.jrc.ec.europa.eu/documents/Metadata/INSPIRE_MD_IR_and_ISO_v1_2_20100616.pdf, [Accessed 23rd March 2011].

JISC (2011), JISC Geospatial Working Group – Terms of Reference [online] Available from: http://www.jisc.ac.uk/aboutus/howjiscworks/committees/workinggroups/geospatial.aspx [Accessed 21st November 2011].

JISC (2011b), JISC Digital Infrastructure Team - Grant Funding Call [online] Available from: http://infteam.jiscinvolve.org/wp/2010/09/10/jisc-grant-funding-call/ [Accessed 21st November 2011].

Kalantari, M., Olfat, H., Rajabifard, A., (2010), Automatic spatial metadata enrichment: reducing metadata creation burden through spatial folksonomies, in GSDI 12 World Conference: Realising Spatially Enabled Societies, Singapore.

Kim, T., (1999), Metadata for geo-spatial data sharing: a comparative analysis, The Annals of Regional Science 33: 171-181.

Koukoletsos, T., Haklay, M., and Ellul, C., (2011), An automated method to assess Data Completeness and Positional Accuracy of OpenStreetMap in Proceedings of the GIS Research UK 19th Annual Conference GISRUK 2011, University of Portsmouth, Portsmouth.

Longley, P.A., Goodchild, M.F., Maguire, D.J., and Rhind, D.W., (2011), Geographical Information Systems and Science, Third Edition, Hoboken, NJ, Wiley.

Manso-Callejo, M.A., Wachowicz, M., and Bernabé-Poveda, A., (2009), Automatic Metadata Creation for Supporting Interoperability Levels of Spatial Data Infrastructures, in GSDI 11 World Conference: Spatial Data Infrastructure Convergence: Building SDI Bridges to Address Global Challenges, Rotterdam, The Netherlands, June 15-19, 2009.

Moellering, H., (2005), *World Spatial Metadata Standards*, Elsevier, Oxford.

Olfat, H., Rajabifard, A., and Kalantari, M., (2010), Automatic Spatial Metadata Update: A new approach in Proceedings of the FIG Congress, Sydney.

Pasca, M., Petriglia, L., Mattioni, F., Torchio,M., and Mariotti, C., (2009), Experiences in the Creation and Updating of INSPIRE Compliant Metadata Catalogue, GSDI11 World Conference: Spatial Data Infrastructure Convergence: Building SDI bridges to address global challenges, Rotterdam, The Netherlands, June 15-19, 2009.

Poore, B., and Wolf, E., (2010), The Metadata Crisis – Can geographic information be made more usable? U.S. Geological Survey Proceedings of the Second Workshop on Usability of Geographic Information, London, March 23, 2010. Available from: http://www.virart.nottingham.ac.uk/GI%20Usability/Workshop%20papers%20pdfs/ [Accessed March 1, 2011].

Rajabifard, A., Kalantari,M., and Binns,A., (2009), SDI and Metadata Entry and Updating Tools" in B. van Loenen, J.W.J. Besemer and J.A. Zevenbergen (Eds.) SDI convergence: Research, Emerging Trends and Critical Assessment, Netherlands Geodetic Commission, Delft.

Sboui, T., Salehi, M., and Bédard,Y., (2009), Towards a Quantitative Evaluation of Geospatial Metadata Quality in the Context of Semantic Interoperability, 6th International Symposium on Spatial Data Quality, St. John's, Newfoundland, Canada, July 6-8, 2009.

Schuurman, N., (2009), The new Brave New World: geography, GIS, and the emergence of ubiquitous mapping and data, *Environment* and Planning D: Society and Space, 27: 571-580.

Servigne, S., Lesage, N., and Libourel, T., (2006), Quality Components, Standards and Metadata, in Devillers R. and Jeansoulin R., editors, Fundamentals of Spatial Data Quality, pp179-208. Fundamentals of Spatial Data Quality, 179-208.

Sboui, T., Salehi, M., and Bedard, Y., (2009), Towards a Quantitative Evaluation of Geospatial Metadata Quality in the Context of Semantic Interoperability, Proceedings of the 6th International Symposium on Spatial Data Quality.

SECOA (2011a), SECOA Project Description [online] Available from: http://projectsecoa.eu/ [Accessed 23 March 2011].

SECOA (2011b), SECOA Case Studies [online] Available from: http://projectsecoa.eu/index.php?option=com_content&view=article&id=50&Itemid=73 [Accessed 23 March 2011].

van Oort, P., (2005), Spatial Data Quality: From Description to Application, PhD thesis, Wageningen University, The Netherlands.

Walker, R., (2009), UK GEMINI Standard - A UK Metadata Standard for discovery of geographic data resources, Association for Geographic Information.

West, L. and Hess,T., (2002), Metadata as a knowledge management tool: supporting intelligent agent and end user access to spatial data, Decision Support Systems, 32: 247-264.