STATISTICAL COMMISSION and Working Paper No. 15
ECONOMIC COMMISSION FOR EUROPE English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Methodological Issues Involving the Integration of Statistics and Geography
(Neuchâtel, Switzerland, 10-12 April 2000)

Topic (i): Policy and organisational aspects in GIS and statistics

USING METADATA IN MULTI-STEP PREPROCESSING AND LONG-TERM MONITORING

Submitted by the Swiss Federal Institute for Snow, Forest and Landscape Research[1]

Contributed paper

I. OVERVIEW

1. The project for a comprehensive storage and retrieval system for environmental data at the Swiss Federal Institute for Snow, Forest and Landscape Research (WSL) has two main objectives:

> (i) to handle temporal and spatial information in a central data repository accessible over a network and provide database functionality;
> (ii) to preserve the semantic context for sensible data generation steps in order to handle data for long-term monitoring projects.

2. The issue of this paper is limited to the second project task and tries to show links between our findings concerning the data model (conceptual level) and metadata approaches in statistics. Our point of departure was the neccessity to handle what we might call 'sensible scientific workflows', e.g. the processing of photogrammetrical data (Figure 1, the English version will be provided at the Work Session).

3. In order to preserve the context of each data generation step (in classical workflow management called "task"), we developed the 'Process Model' (PM). The PM is a datamodel comprising data structure, update and retrieval operations and integrity constraints, essentially based on ORDBMS functionality and which uses a special combination of metadata in a specific way - specific in that metadata are not used to describe complex data objects at the instance level, but to restrict the metadata content exclusively for a comprehensive description of each data generation step. Descriptive information of the data itself is, in this model, treated as inferable information, based on system data dictionaries, code tables and the above-mentioned process metadata through several data generation steps. The goal of this approach is to keep the underlying data model clean of application and domain specific metadata and thus extensible for different application domains without changing the underlying database scheme. The same data model can therefore be used for a wide range of scientific processing tasks.

4. Method: The paper starts with a comparison of the PM towards similar information handling concepts. This comparison is followed by an overview of the basic features of the data model, the relevant metadata contents and its entity model. An example of an iterative collection of metadata shows how data descriptions can be deduced from the process metadata. This exemplifies at the same time that data

[1] Prepared by Nick Baumberger and Martin Hägeli.

descriptions are implicitly contained within the metadata and hence must not be modelled explicitly. Finally, this paper shows how a simple statistical workflow can be implemented in the process model.

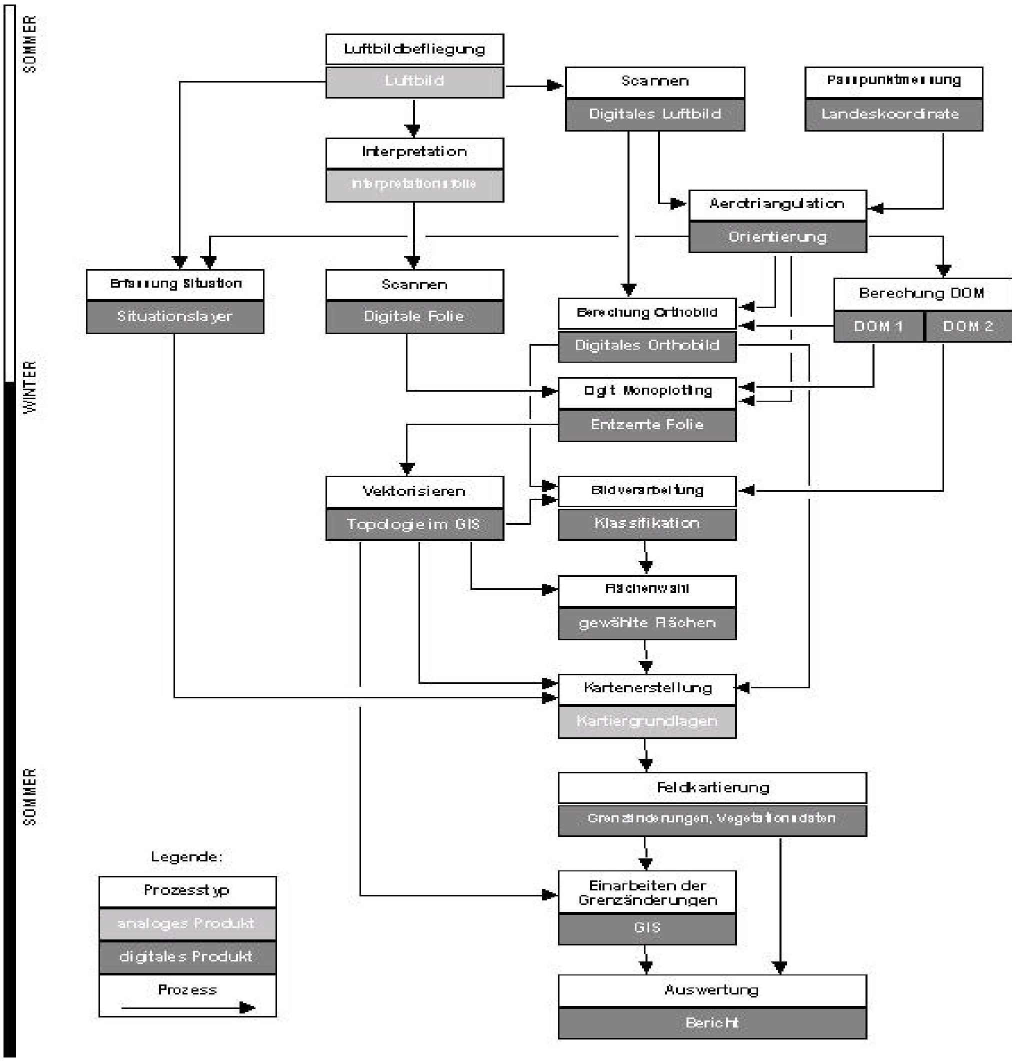

Figure 1: Example of a scientific workflow: photogrammetrical preprocessing

II. SIMILARITIES AND DIFFERENCES TO OTHER INFORMATION SYSTEMS SOLUTIONS

5. Workflow Management Systems (WFMS): WFMS are used to actively control and monitor repetitive tasks. In the typical conceptualization of workflows, the focal point is the action, i.e. the processes that take place during workflow execution. Workflows are considered as transactions, with the information they manipulate playing a subordinate role. This transactional view of workflows leads to an architecture followed by the majority of existing WFMS. A dedicated workflow specific software runs on top of a DBMS, i.e. process management is separated from data management. [Ailamaki et al., 1998].

6. The interaction between user and system, in the presented PM, is not controlled by an active ‘process scheduler’. The PM functionality is only concerned in constraining the input data in order to ensure process logic and completeness of the metadata.

7. Document Management Systems (DMS): Document Management requires central repositories, control over access to information, consistent use of document formats and processes of workflow. In comparison with DMS, the heterogeneity of data types (domains) handled by the PM is much higher. Therefore, a lot of DMS functionality is either not available or simply doesn’t make sense for the PM. The core functionality is restricted to data structure, conversion, access, standardised metadata content and

automated import of documents. Many typical DMS functions such as document delivery, full-text retrieval, document discovery, import, removal and document flow are set aside completely or, in the case of document description [2], are only partly discussed [Wilkinson et al., 1998].

8. Process Synchronisation in WFM: As described in Alonso, "[.] the idea of such a system is to provide a way to use existing tools, linked by a network as building blocks of higher level system in which the process acts as the blueprint for control and data flow". Typical examples are business processes in which several office tools such as spread-sheets, text editors, database and human decisions are combined into a higher level entity by coding the business logic within the flow of control and data of the process [3] [Alonso 1996, 1997].

9. There are many common aspects between synchronizing distributed applications over a network and describing the data processing context in scientific workflows, but the basic idea of Process Synchronization is to couple the synchronisation mechanism with the scheduler algorithms of the WFMS in such a way that activities are only started when consistency can be guaranteed. This is primarily a software engineering task whereas the main task of the PM is restricted to determining the content and structure of metadata to guarantee the reproducibility of every single step in the scientific work chain.

II. PROCESS MODEL (PM)

II.1 The 'process' of data generation

10. The smallest unit of the PM described in this paper is the 'process'. The process is a single act in which new information is generated, in other words, a data generation step. Each single process is considered as a transformation process. The transformation is not constrained to computerized information handling. Non-digital information such as observations or estimations in the field are considered as data generation steps, as are automated precipitation or sunlight measurements. What all these processes have in common is the use of specific tools and know-how in order to produce new information. We might put it in a more abstract sense; that in all efforts to measure, analyse and interpret the world around us, we proceed with the same principle: transform input information in a specific way into output-information. This is just what a single process ought to model.

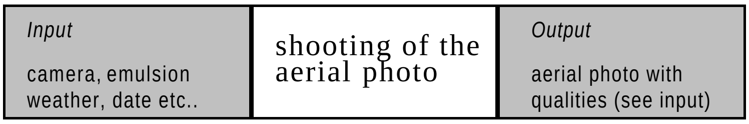

Figure 2: Input-Output-Model: Shooting of an aerial photo

II.2 Process chain: chaining of processes

11. In scientific workflows especially, information is developed through many information processing steps. In order to model process sequences and versioning (forking of workflows), explicit references to predecessor processes are stored as part of the metadata. The links to the direct predecessor in the data generation chain serve also as plausibility checks. Through these backward or 'upward' directed links, the history of every item of processed data can be reproduced.

[2] Document Description Languages (DDL): ASCII, Unicode, RTF, TeX, LaTeX, SGML, HTML, XML, Postscript, pdf. [Croft, 1998, 25ff]

[3] In the context of transactional process management: „[.] a process can be seen as a description of an arbitrary sequence of application invocations along with the data flow between these applications. As such, the process acts as a meta-program governing the interactions among existing applications. Each step within a process is an activity, which represents invocations of external applications." [Alonso, 1997, 2]

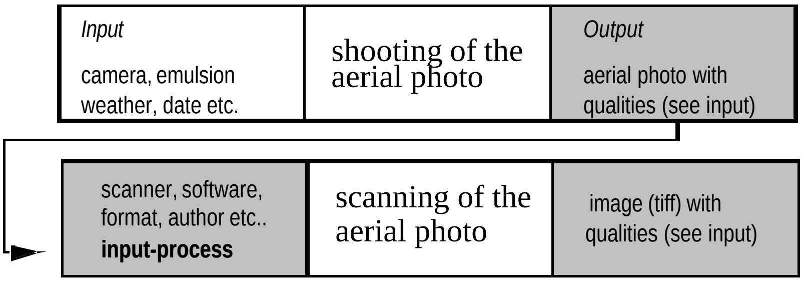

Figure 3: Chaining information as part of metadata

III. METADATA IN THE CONTEXT OF PROCESSES

III.1 Metadata content: metadata ensuring reproducibility of the workflow

12. Metadata are exclusively used for the description of the data generation step and not for the description of the generated data itself [4]. The following information is considered essential to guarantee (interpersonal) reproducibility of a data generation step:

- technical instruments, tools (specially mechanical, optical tools), hardware, software;
- operating systems;
- actor’s information, workspace, contact information;
- detailed information subdividing the single process into several phases (i.e. for accounting purposes);
- comments, special remarks;
- source-codes of scripts or applications, executables;
- program calls, parameters and other instructions;
- format informations.

13. This information is stored in a codified form for each generated data item, i.e at record level. It is not necessary to store further detailed technical descriptions of instruments, software, hardware, os, etc. at this level. They can be stored at schema level as look-up information and referenced when needed (Table 1).

14. In this model, quality information of GIS data is not modelled as a feature of a single or a set of geometrical objects, but is deduced from the original input data, e.g. the cartographic draft or the technical specifications of the digitizing table or the experience of the digitizing person. All this information represents input metadata. The total set of metadata concerning a data object, generated through several preprocessing steps, is therefore not attached directly, but must be iteratively retrieved from each processing step (‘process’) in the chain. The presentation and formatting of the retrieved metadata is left to the requirements of the output standard. Presentation and representation of metadata are therefore separated [5].

III.2 Metadata entities

15. To preserve thematic extensibility, only the greatest common denominator of meta information, as it is shared in different application domains, is modelled as structured (normalized) attributes. The values are codified and described in simple code tables. Additional, more detailed information related to the metadata codes is stored in separate, so-called extended code tables (Table 2). All application specific information is held in unstructured fields we call ‘metalobs’. Separating in this way the commonly used metadata and the type specific metadata, it is possible to extend the thematic content without changing the underlying database scheme, simply by creating additional metalob tables. There are four metalob tables planned to hold unstructured metadata: commands & parameters, binaries,

[4] Further information on metadata, see FDGC (1997), Jefferey (1998), GSF (1999), Esri (1995).

[5] The external and conceptual level of the metadata model are independent of each other (logical data independence).

comments and unstructured quality information [6]. In pursuing the principle of minimizing the amount of structured information, we minimize the risk of scheme adoptions. Actually, there are 28 metadata attributes whereas 2 fields are used for chaining, 2 fields are system dependent, 9 fields contain key information and the remaining 15 fields contain the metadata codes.

Table 1: Structured metadata

Column	Use
procno_in extern_in_id	Chaining
Tool Hardware Software Os Proclookup Autor ...	Codified structured metadata

Table 2: Extended code table (camera specifications)

Column	Use
Code	Coding
focal length principle point fiducial marks	Description of camera
scann distortion scann frequence	Description of scanner

III.3 ERD: processdata and metadata

16. The complete set of process information consists of the generated data itself, called ‘processdata’ and the structured and unstructured metadata. Possible processdata can be text documents, geometrical boundaries, raster data, statistic plots, etc. According to its domain this data is stored in different tables, e.g. spatial information can be put together in a layer table, documents in pdf format can be stored in one ‘docu’ table, or rasterdata can be held in a separate image table together with other tiff data [7].

V. STEPWISE METADATA RETRIEVAL AND FORMATTING

17. An image example shows how metadata can be retrieved, filtered and formatted by navigating stepwise through the process chain (Table 2, Figure 3). In the first step, metadata concerning the shooting of the picture is retrieved, e.g. technical details about the camera such as focal length, distortion, emulsion type or flight altitude. In a second step the metadata concerning the scanning of the analog aerial photo is described with the parameter scann resolution. The total of this information can be filtered and formatted as shown in Table 3, which is a normal form table with attribute categories as records and ‘/’-delimited values in one column.

[6] Before uploading, this ‘unstructured’ information can be formatted according the needs of data retrieval (e.g. XML style formatting of ASCII documents).

[7] On the conceptual design level, there exists a generalisation-specialisation relation between the meta- and the process data. This relation is modelled in an additional lookup table.

Figure 4: ERD of the processdata and metadata

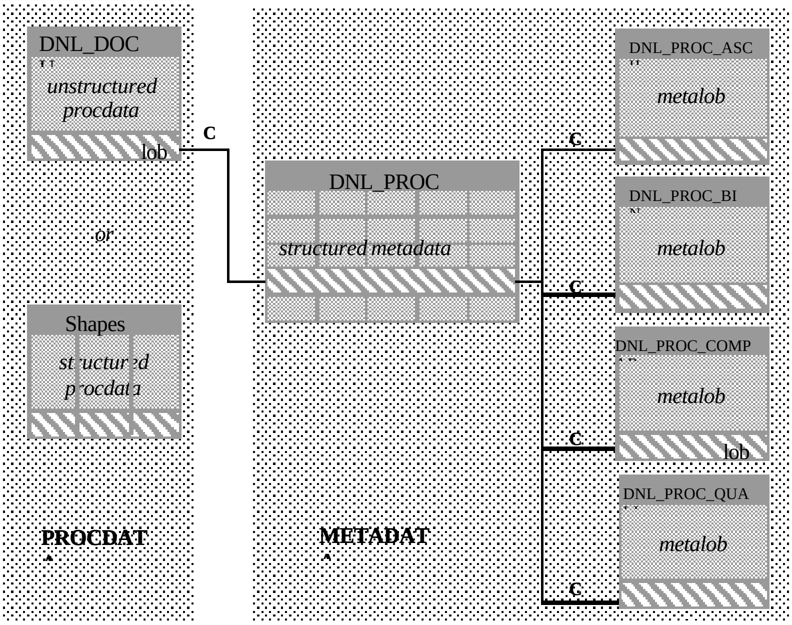

Table 3: Presentation format of descriptive metadata (shooting and scanning of an aerial photo)

Metadata category	N1NF formatted metadata values
tool	camera RC70-Leica / Horizon-Agfa
hardware	- / -
software	- / Fotolook
os	- / MacOS8
dataformat	- / tiff
fileformat	- / tif-photoshop
proclookup	- / -
autor	Douglas / Wyder
work_place	KSL / WSL
source_ref	- / -
source_orient	- / -
order_no	1200-88787-0 / -
resolution	- / 600 dpi
focal length	70 mm / -
distortion	- / 0.001

V. STATISTICAL WORKFLOWS IN THE PROCESS MODEL

18. In general, statistical information processing proceeds from microdata (tuples, observation data) to macrodata by applying summary functions to microdata. The macrodata consists of the statistical summary values. Tuple retrieval is of little interest as its focus is on summarizing and filtering average characteristics. Froeschl writes: “On its passage from microdata to macrodata, statistical information processing runs through a sequence of stages [.]. It is often macrodata that is the real point of departure” (Table 4). It is argued that summary attributes have a semantic quality entirely different from category attributes which do not represent individual observations but collections sharing the same patterns of selected observable features. In order to

ensure that the further processing of macrodata is adopted appropriately, a knowledge of the data semantics and the way the macrodata has been generated is necessary [Froeschl, 1997, 9,26ff].

Table 4: Data processing stages vs. data levels [Froeschl, 1997, 27]

Step	Processing stage	Data level	Dimensionality
1	raw dataset	micro	-
2	edited, imputed, cleaned dataset	micro	-
3	(optional) anonymized dataset	micro	-
4	structure conversion	pre-macro	high
5	elementary counting	pre-macro	high
6	(optional) anonymizing	pre-macro	high
7	weighting, estimation	lower-level-macro	high
8	(optional anonymizing	lower-level-macro	high
9	(optional) pre-aggregation	lower-level-macro (summary set)	medium
10	final aggregation (stockpile)	higher-level-macro	low
11	post-processing	higher-level-macro	low

VI. STATISTICAL METADATA MODELLING APPROACHES

19. Statistical metadata is modelled in order to attain a coherent semantical model so that the context of each processing step is preserved. A common characteristic of the discussed metadata modelling approaches is to establish semantics with relations and references between data objects. In this way, a meta structure is built around the data. In the so-called ‘Subject Model’, Chen presents this referential information as a graph, displaying a network of relations between aggregated values at different processing stages, subselections and raw data.

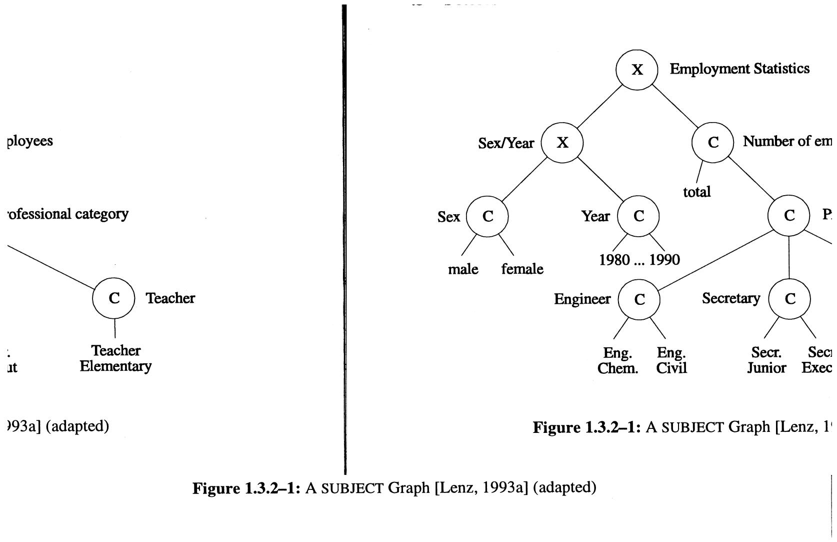

Figure 5: A subject graph according to Chen and Shoshani (1981) [Froeschl, 1997]

20. Very similarly, in the process model, we understand processes as a network of sequential and parallel processing steps that can be visualized by graphs. How such a workflow can be implemented is shown in an evaluation of spatially distributed findings of lichen. The process chain includes: 1) selection of relevant data, 2) calculation of some statistical aggregates and 3) interpretation of the results in a report. The example can be found under the lichen projects at http:\\www.wsl.ch.

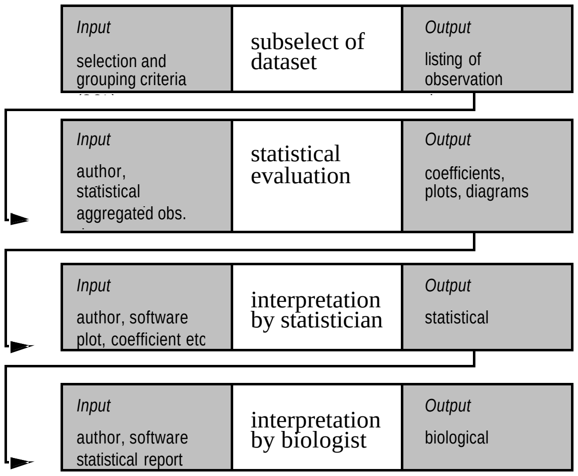

Figure 6: Simple statistical workflow in the PM

REFERENCES

Ailamaki A., Ioannidis Y.E., Livny M., 1998. Scientific Workflow Management by Database Management. IEEE, 190-199, IEEE Computer Society, Danvers MA

Alonso G., 1997. Processes + Transactions = Distributed Applications. Position Paper HPTS'97, Institute of Information Systems, Swiss Federal Institute of Technology (ETH), Zürich

Alonso G., Agrawal D., El Abbadi A., 1996. Process Synchronization in Workflow Management Systems 8the IEEE Symposium on Parallel and Distributed Processing (SODS'97). New Orleans, Louisiana - October 23 - 26, 1996.

ESRI, 1995. Metadata Management in GIS. ESRI White Paper Series, Redlands CA

Federal Geographic Data Committee (FDGC), 1997. Content Standard for Digital Geographic metadata. Washington DC, www.fdgc.gov

Froeschl K.A., 1997. metadata Management in Statistical Information Processing. Springer Computer Science, Wien.

Ginzler Ch., Hägeli M., De Laporte K., Mauser H., Thee P., 1999. Wie kommt das Moor ins GIS ? - Der Einsatz der Photogrammetrie bei der Wirkungskontrolle Moorbiotope Schweiz. Vermessung, Photogrammetrie, Kulturtechnik, No. 9/1999, 483-489, Hrsg. Schweizerischer Verein für Vermessung und Kulturtechnik (SVVK), Solothurn.

GSF - Forschungszentrum für Umwelt und Gesundheit, 1997. UFIS Concept for Metainformation on Data, www.gsf.de/UFIS/ufis/ufis_proj.html

Jefferey K.G., 1998. Metadata: An Overview and some issues. Advanced Database and Metadata, ERCIM Online News No.35, www.ercim.org

Wilkinson R., Arnold-Moore T. et al., 1998. Document Computing: Technologies for Managing Electronic Document Collections, The Kluwer international series on information retrieval, Kluwer Academic Publishers, Boston