Unit#4c

Data Protection

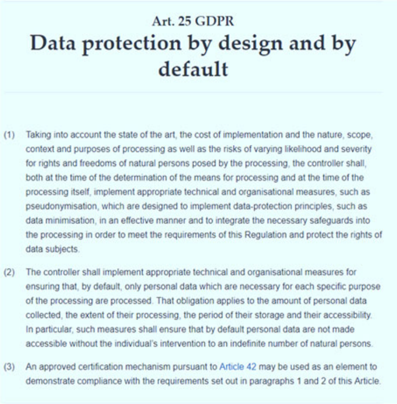
MIS5214

Agenda

- Data protection by design
- System Security Plan
  - Security control inheritance
  - Team project SSP review and discussion

Data security by design and default…

Data protection capabilities must work from beginning to end of data processing to enable protection of individuals' personal data by default



Art. 25 GDPR

Data protection by design and by default

(1) Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

(2) The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

(3) An approved certification mechanism pursuant to Article 42 may be used as an element to demonstrate compliance with the requirements set out in paragraphs 1 and 2 of this Article.

Danezis, G. et al. (2014) "Privacy and Data Protection by Design",
European Union Agency for Network and Information Security (ENISA)

D' Acquisto, G. et al. (2015) "Privacy by design in big data",
European Union Agency for Network and Information Security (ENISA)

Key General Data Protection Regulation (GDPR) requirements:

1. **Collection** of personal data is **fully avoided or minimized** at the earliest stage of processing
2. Data subjects give **specific**, **informed** **and** **explicit** **consent** to the processing of their data
3. Data subjects have **right to access, review and rectify** their personal data
4. Data subjects have the **right to withdraw given consent** with effect for the future and
   - Block access
   - Constrain processing and use
   - Erase their personal data
5. Personal **data obtained for one purpose must not be processed for other purposes** not compatible with the original purpose

Achieving “Privacy by Design” is difficult

Privacy is a complex, multifaceted and contextual notion

Not the primary requirement of an information system

May come into conflict with other requirements

“…privacy and data protection features are… ignored by traditional engineering approaches when implementing desired functionality.

- *This ignorance is caused by limitations of awareness and understanding of developers and data controllers as well as lacking tools to realize privacy by design”*

Danezis, G. et al. (2014) “Privacy and Data Protection by Design”,
European Union Agency for Network and Information Security (ENISA)

Privacy and Data Protection by Design

“Although the concept has found its way into legislation as the… European General Data Protection Regulation, **its concrete implementation remains un-clear at the present moment**”

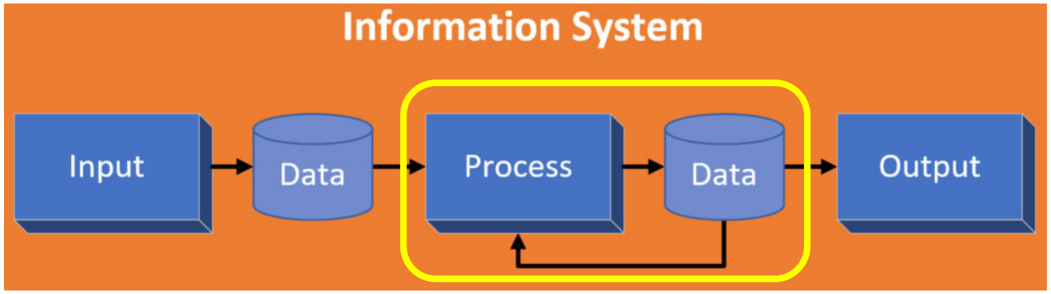
Danezis, G. et al. (2014) “Privacy and Data Protection by Design”,
European Union Agency for Network and Information Security (ENISA)

Some challenging data protection requirements may be solved with techniques presented here…

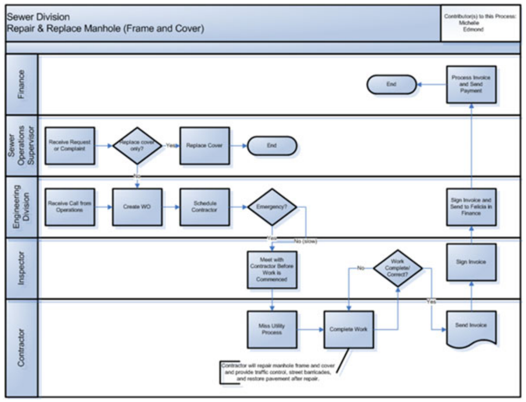
1. **Collection** of personal data is **fully avoided or minimized** at the earliest stage of processing
2. Data subjects give **specific**, **informed and explicit consent** to the processing of their data
3. Data subjects have **right to access, review and rectify** their personal data
4. Data subjects have the **right to withdraw given consent** with effect for the future and
   - Block access
   - Constrain processing and use
   - Erase their personal data
5. Personal **data obtained for one purpose must not be processed for other purposes** not compatible with the original purpose

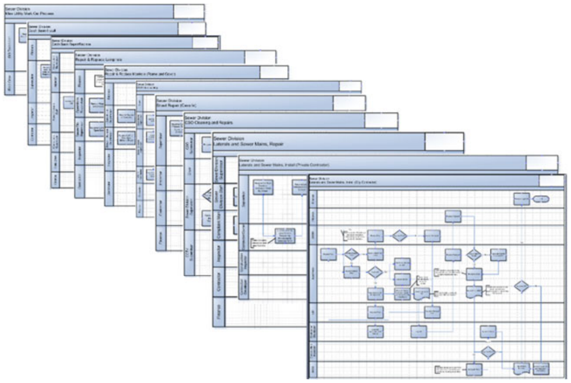
As a practical matter…

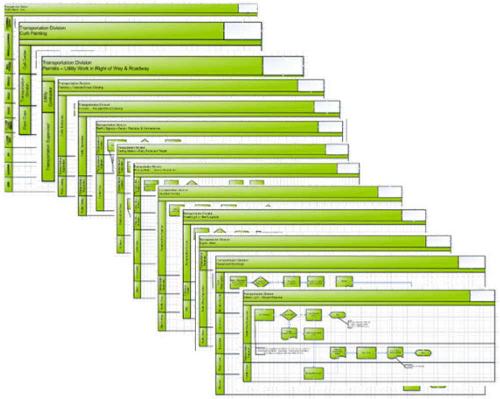
Data within information systems are often stored and organized as datasets within files and/or databases…



Regardless of application, there is reliance on data processing workflows to produce and use information

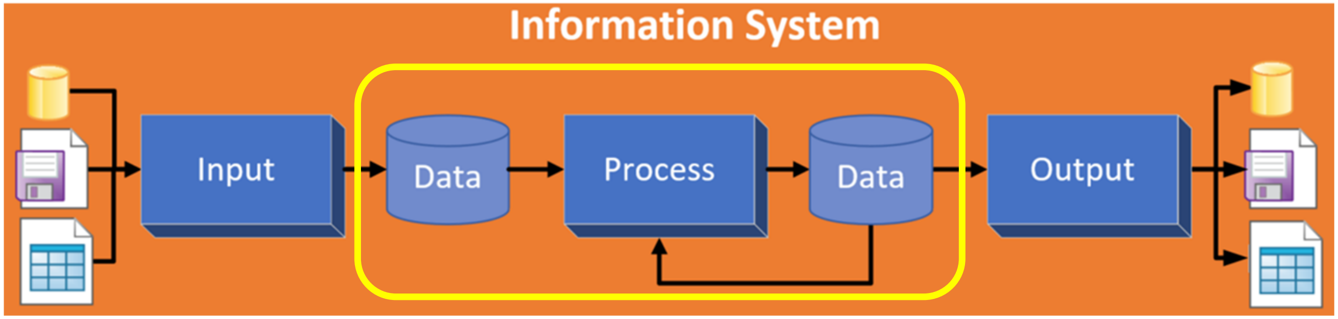

Sewer Division
Repair & Replace Manhole (Frame and Cover)
Contributor(s) to this Process: Michelle Edmond
Finance
Sewer Operations Supervisor
Engineering Division
Inspector
Contractor
End
Process Invoice and Send Payment
Receive Request or Complaint
Replace cover only?
Yes
Replace Cover
End
No
Receive Call from Operations
Create WO
Schedule Contractor
Emergency?
Sign Invoice and Send to Felicia in Finance
Yes
No (slow)
Meet with Contractor Before Work is Commenced
No
Work Complete/ Correct?
Sign Invoice
Yes
Miss Utility Process
Complete Work
Send Invoice
Contractor will repair manhole frame and cover and provide traffic control, street barricades, and restore pavement after repair.



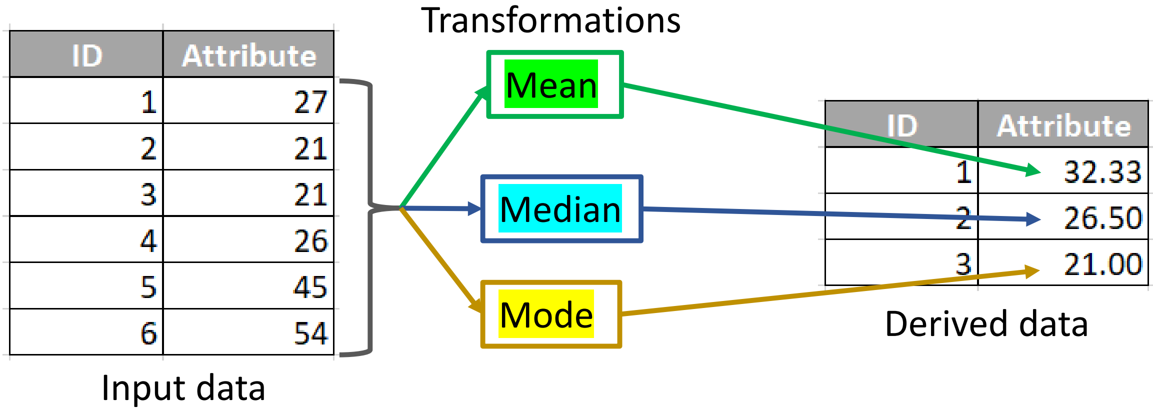


Data processing often transforms existing data into new data, which is a double-edged sword…

- *The resulting database may have more information than the older version*



- *The **meaning** of the new information, however, **is exogenous and not found in the data itself***



ID	Attribute
1	27
2	21
3	21
4	26
5	45
6	54

Input data

ID	Attribute
1	32.33
2	26.50
3	21.00

Derived data

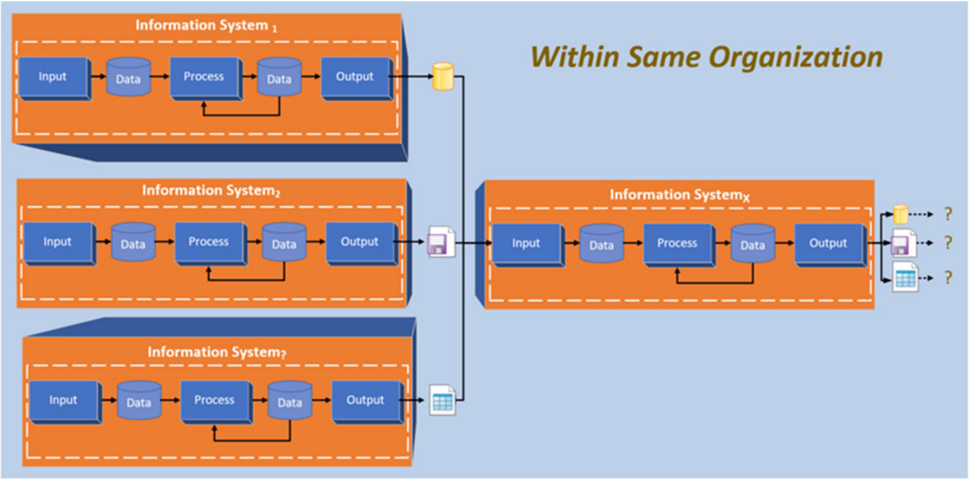
Evaluating & judging data's "fitness for use"

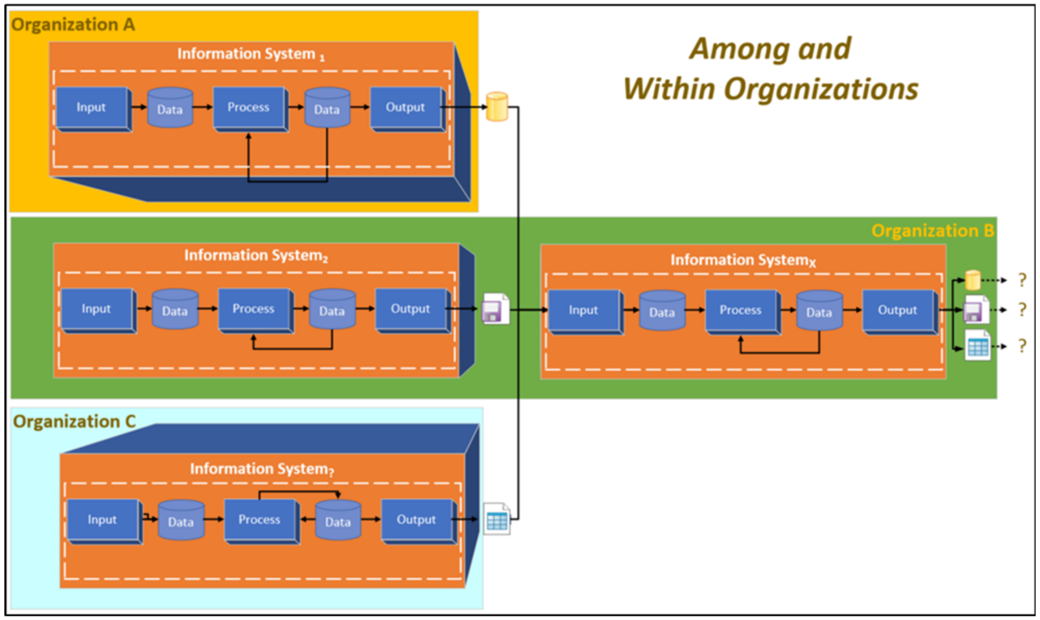
- **Is not the responsibility of the producer**
- **Is the responsibility of the user …and IT Auditor**

Data produced for one purpose is often used to serve other purposes

Data producers should provide information about data that permit informed determinations of fitness for use

Datasets are often exchanged without information needed to determine their fitness for use…





Provenance

The Bridge at Villeneuve-la-Garenne
1872
by Alfred Sisley British

Provenance traces back to 1294 in Old French as a derivative of the Latin *provenire*

- *To come from, to be due to, be the result of*

In the art domain provenance entails an artifact's complete ownership history

Provenance refers to the **origin, history, and journey of data** as it moves through different systems, transformations, and users. It ensures **transparency, traceability, and accountability** by documenting how data was created, modified, and shared.

There is an established research process for obtaining an artifact's trusted provenance

- *The information is highly valued, particularly to authenticate real versus fraudulent works*

"Provenance" is now increasingly used in a broad range of fields with various degrees of conflation of two closely related but distinct concepts of trust and metadata

Tullis, J.A. et al., 2016, "Geoprocessing, Workflows, and Provenance", in Remote Sensing Handbook: Remotely Sensed Data Characterization, Classification, and Accuracies, edited by P. Thenkabail, Vol. 1., pp. 401-422, Boca Raton, FL: CRC Press.

Provenance

W3C Provenance Incubator Group's definition of provenance (in a web resource context):

- Provenance is a record that describes entities and processes involved in producing and delivering or influencing a resource
- Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility
- Provenance assertions are contextual metadata that can become important records with their own provenance

https://www.w3.org/TR/prov-primer/

Why is Provenance Important?

- **Transparency**: Helps understand the source and trustworthiness of data.
- **Accountability**: Supports compliance with regulations like **GDPR, HIPAA, and PCI DSS**.
- **Data Integrity**: Ensures that data is accurate, unaltered, and verifiable.
- **Security**: Helps detect unauthorized modifications and data leaks.
- **Audit & Compliance**: Provides a detailed history of data usage for regulatory audits.

Data Lineage

What is Data Lineage?

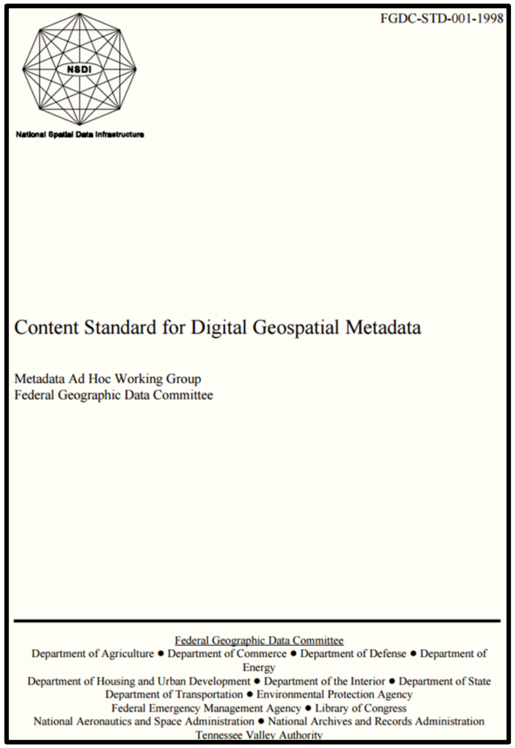
Data lineage refers to the process of tracking **the flow, transformation, and dependencies of data** across different systems and processes. It provides a **visual and structured history** of how data moves through an organization, from its **origin (source)** to its **final destination**

Why is Data Lineage Important?

- **Regulatory Compliance** (GDPR, HIPAA, CCPA): Helps organizations track and verify how personal data is handled.
- **Data Governance**: Ensures accuracy, consistency, and quality of data across the enterprise.
- **Security & Risk Management**: Detects unauthorized modifications or data breaches.
- **Impact Analysis**: Helps understand how changes in data sources affect downstream systems.
- **Debugging & Troubleshooting**: Identifies data errors and inconsistencies.

Early metadata standards for documenting lineage of data produced with Geographic Information Systems

Metadata is "data about data"—it provides **descriptive, structural, and administrative** information about a file, document, image, or dataset. It helps users **organize, find, manage, and understand** data efficiently.

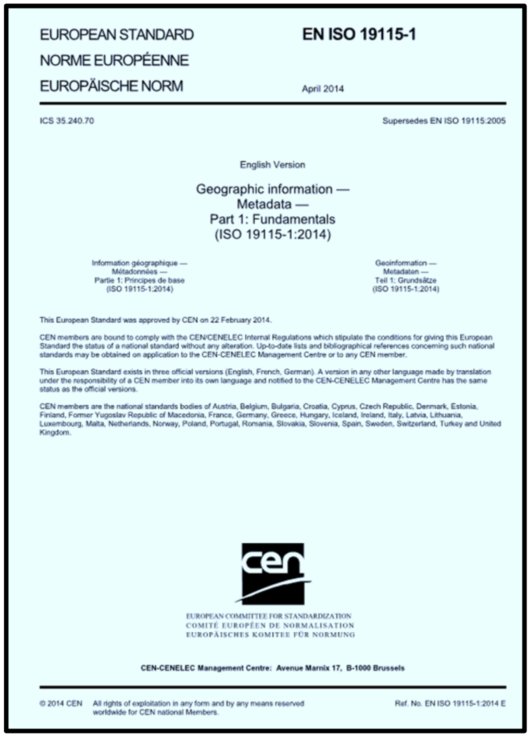


FGDC-STD-001-1998

Content Standard for Digital Geospatial Metadata

Metadata Ad Hoc Working Group
Federal Geographic Data Committee

Federal Geographic Data Committee
Department of Agriculture ● Department of Commerce ● Department of Defense ● Department of Energy
Department of Housing and Urban Development ● Department of the Interior ● Department of State
Department of Transportation ● Environmental Protection Agency
Federal Emergency Management Agency ● Library of Congress
National Aeronautics and Space Administration ● National Archives and Records Administration
Tennessee Valley Authority



EUROPEAN STANDARD
NORME EUROPÉENNE
EUROPÄISCHE NORM

EN ISO 19115-1

April 2014

ICS 35.240.70

Supersedes EN ISO 19115:2005

English Version

Geographic information —
Metadata —
Part 1: Fundamentals
(ISO 19115-1:2014)

Information géographique —
Métadonnées —
Partie 1: Principes de base
(ISO 19115-1:2014)

Geoinformation —
Metadaten —
Teil 1: Grundsätze
(ISO 19115-1:2014)

This European Standard was approved by CEN on 22 February 2014.

CEN members are bound to comply with the CEN/CENELEC Internal Regulations which stipulate the conditions for giving this European Standard the status of a national standard without any alteration. Up-to-date lists and bibliographical references concerning such national standards may be obtained on application to the CEN-CENELEC Management Centre or to any CEN member.

This European Standard exists in three official versions (English, French, German). A version in any other language made by translation under the responsibility of a CEN member into its own language and notified to the CEN-CENELEC Management Centre has the same status as the official versions.

CEN members are the national standards bodies of Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Former Yugoslav Republic of Macedonia, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey and United Kingdom.

EUROPEAN COMMITTEE FOR STANDARDIZATION
COMITÉ EUROPÉEN DE NORMALISATION
EUROPÄISCHES KOMITEE FÜR NORMUNG

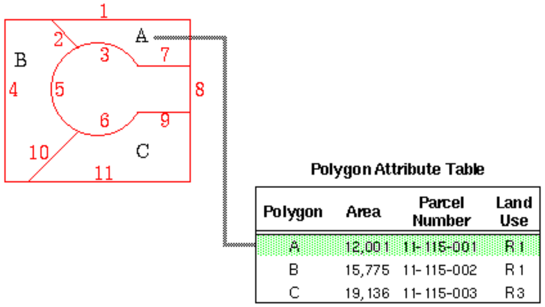
CEN-CENELEC Management Centre: Avenue Marnix 17, B-1000 Brussels

© 2014 CEN All rights of exploitation in any form and by any means reserved worldwide for CEN national Members.

Ref. No. EN ISO 19115-1:2014 E

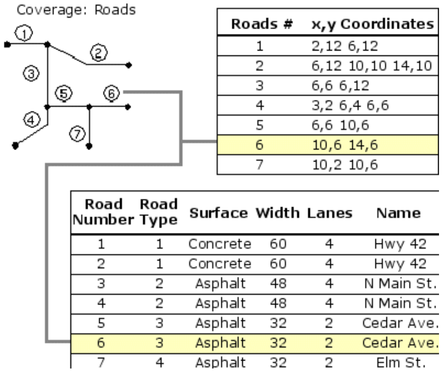
Geographic Information System (GIS)

- Provides similar data import, query, manipulation, analysis (e.g. statistics), reformat, display/visualization, output and report capabilities as other information systems
- Also organize their data in
  - Data base management systems
  - File systems



Polygon Attribute Table

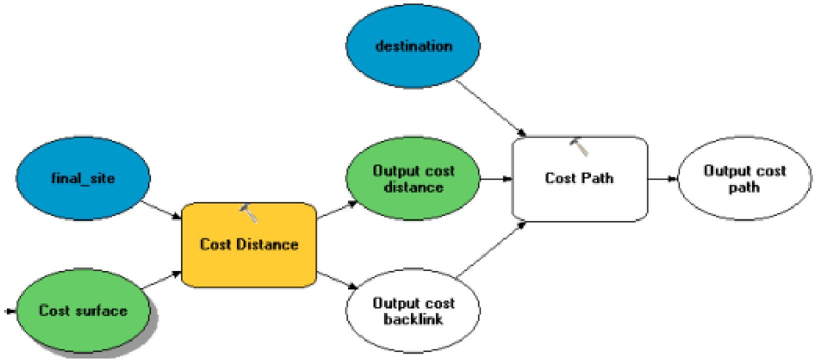
Polygon	Area	Parcel Number	Land Use
A	12,001	11-115-001	R1
B	15,775	11-115-002	R1
C	19,136	11-115-003	R3

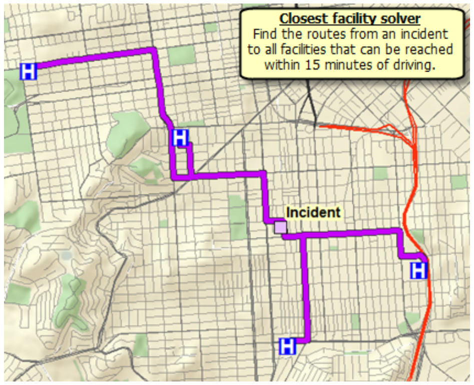

Coverage: Roads

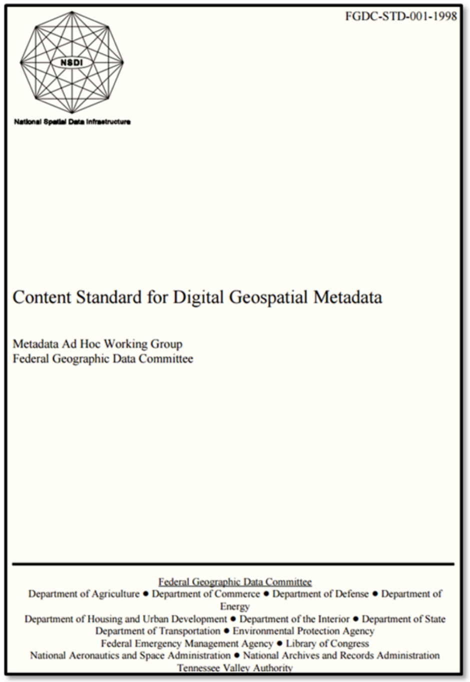
Roads #	x,y Coordinates
1	2,12 6,12
2	6,12 10,10 14,10
3	6,6 6,12
4	3,2 6,4 6,6
5	6,6 10,6
6	10,6 14,6
7	10,2 10,6

Road Number	Road Type	Surface	Width	Lanes	Name
1	1	Concrete	60	4	Hwy 42
2	1	Concrete	60	4	Hwy 42
3	2	Asphalt	48	4	N Main St.
4	2	Asphalt	48	4	N Main St.
5	3	Asphalt	32	2	Cedar Ave.
6	3	Asphalt	32	2	Cedar Ave.
7	4	Asphalt	32	2	Elm St.

- With the addition of spatial analysis and cartographic mapping capabilities




Closest facility solver
Find the routes from an incident to all facilities that can be reached within 15 minutes of driving.

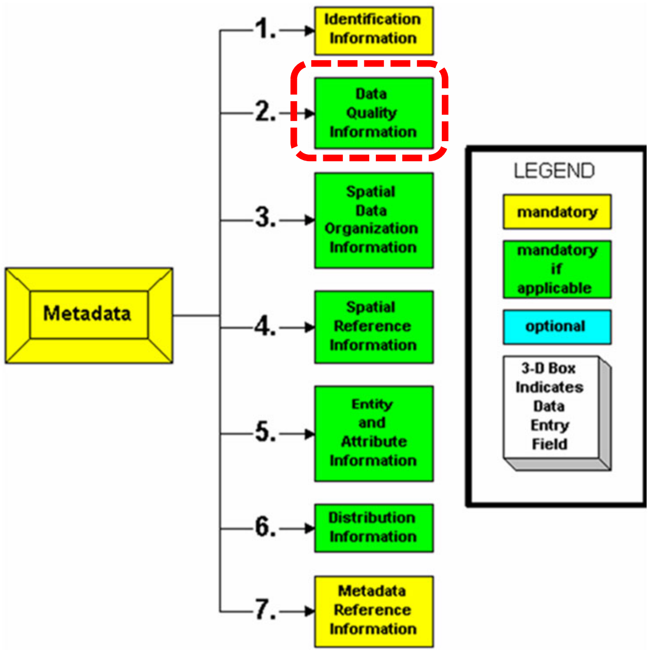


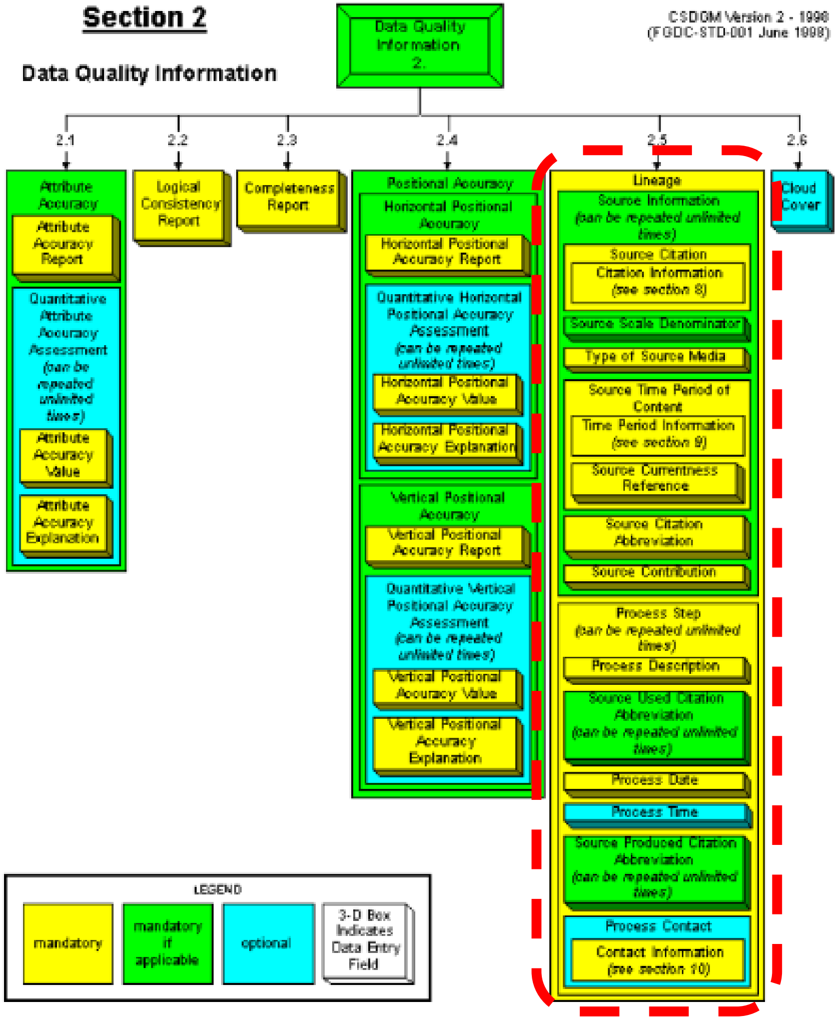
FGDC-STD-001-1998

Content Standard for Digital Geospatial Metadata

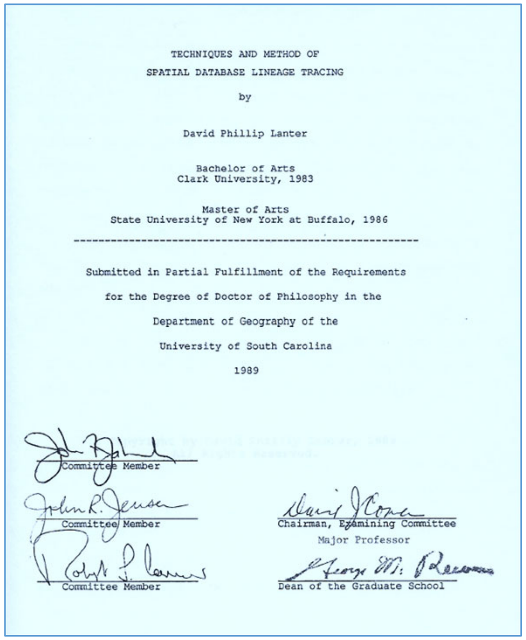
Metadata Ad Hoc Working Group
Federal Geographic Data Committee

Federal Geographic Data Committee
Department of Agriculture • Department of Commerce • Department of Defense • Department of Energy
Department of Housing and Urban Development • Department of the Interior • Department of State
Department of Transportation • Environmental Protection Agency
Federal Emergency Management Agency • Library of Congress
National Aeronautics and Space Administration • National Archives and Records Administration
Tennessee Valley Authority


Metadata
1.
Identification Information
2.
Data Quality Information
3.
Spatial Data Organization Information
4.
Spatial Reference Information
5.
Entity and Attribute Information
6.
Distribution Information
7.
Metadata Reference Information
LEGEND
mandatory
mandatory if applicable
optional
3-D Box Indicates Data Entry Field


Section 2
Data Quality Information
Data Quality Information 2.
CSDGM Version 2 - 1998
(FGDC-STD-001 June 1998)
2.1
Attribute Accuracy
Attribute Accuracy Report
Quantitative Attribute Accuracy Assessment (can be repeated unlimited times)
Attribute Accuracy Value
Attribute Accuracy Explanation
2.2
Logical Consistency Report
2.3
Completeness Report
2.4
Positional Accuracy
Horizontal Positional Accuracy
Horizontal Positional Accuracy Report
Quantitative Horizontal Positional Accuracy Assessment (can be repeated unlimited times)
Horizontal Positional Accuracy Value
Horizontal Positional Accuracy Explanation
Vertical Positional Accuracy
Vertical Positional Accuracy Report
Quantitative Vertical Positional Accuracy Assessment (can be repeated unlimited times)
Vertical Positional Accuracy Value
Vertical Positional Accuracy Explanation
2.5
Lineage
Source Information (can be repeated unlimited times)
Source Citation
Citation Information (see section 8)
Source Scale Denominator
Type of Source Media
Source Time Period of Content
Time Period Information (see section 9)
Source Currentness Reference
Source Citation Abbreviation
Source Contribution
Process Step (can be repeated unlimited times)
Process Description
Source Used Citation Abbreviation (can be repeated unlimited times)
Process Date
Process Time
Source Produced Citation Abbreviation (can be repeated unlimited times)
Process Contact
Contact Information (see section 10)
2.6
Cloud Cover
LEGEND
mandatory
mandatory if applicable
optional
3-D Box Indicates Data Entry Field

The first metadata system focused on GIS data lineage



TECHNIQUES AND METHOD OF
SPATIAL DATABASE LINEAGE TRACING

by

David Phillip Lanter

Bachelor of Arts
Clark University, 1983

Master of Arts
State University of New York at Buffalo, 1986

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in the
Department of Geography of the
University of South Carolina

1989

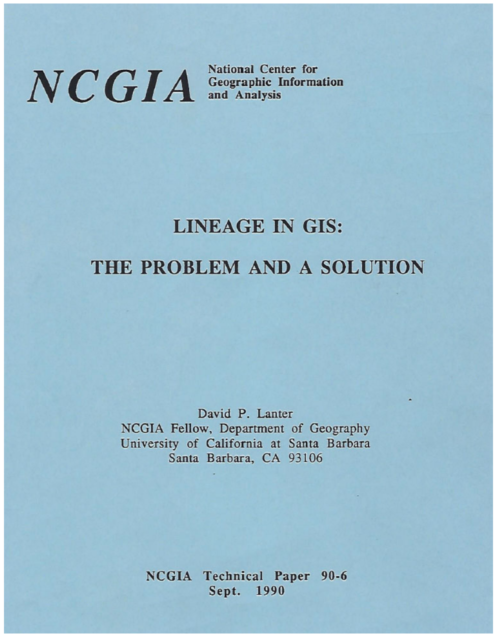
Committee Member

John R. Jensen
Committee Member

Chairman, Examining Committee
Major Professor

Committee Member

Dean of the Graduate School



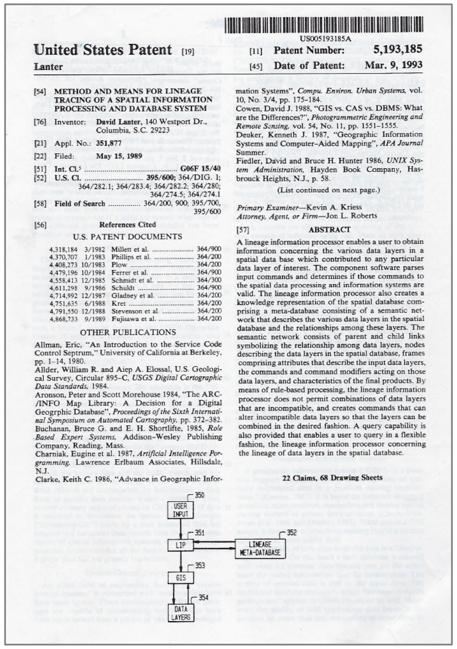
NCGIA National Center for Geographic Information and Analysis

LINEAGE IN GIS:

THE PROBLEM AND A SOLUTION

David P. Lanter
NCGIA Fellow, Department of Geography
University of California at Santa Barbara
Santa Barbara, CA 93106

NCGIA Technical Paper 90-6
Sept. 1990



US005193185A

United States Patent [19]
Lanter

[11] **Patent Number: 5,193,185**
[45] **Date of Patent: Mar. 9, 1993**

[54] **METHOD AND MEANS FOR LINEAGE TRACING OF A SPATIAL INFORMATION PROCESSING AND DATABASE SYSTEM**

[76] Inventor: **David Lanter**, 140 Westport Dr., Columbia, S.C. 29223

[21] Appl. No.: **351,877**

[22] Filed: **May 15, 1989**

[51] **Int. Cl.**5 .... **G06F 15/40**
[52] **U.S. Cl.** **395/600;** 364/DIG. 1; 364/282.1; 364/283.4; 364/282.2; 364/280; 364/274.5; 364/274.1
[58] **Field of Search** 364/200, 900; 395/700, 395/600

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,318,184	3/1982	Millett et al.	364/900
4,370,707	1/1983	Phillips et al.	364/200
4,408,273	10/1983	Plow	364/200
4,479,196	10/1984	Ferrer et al.	364/900
4,558,413	12/1985	Schmidt et al.	364/300
4,611,298	9/1986	Schuldt	364/900
4,714,992	12/1987	Gladney et al.	364/200
4,751,635	6/1988	Kret	364/200
4,791,550	12/1988	Stevenson et al.	364/200
4,868,733	9/1989	Fujisawa et al.	364/200

OTHER PUBLICATIONS

Allman, Eric, "An Introduction to the Service Code Control Septrum," University of California at Berkeley, pp. 1–14, 1980.
Allder, William R. and Aiep A. Elossal, U.S. Geological Survey, Circular 895–C, *USGS Digital Cartographic Data Standards,* 1984.
Aronson, Peter and Scott Morehouse 1984, "The ARC-/INFO Map Library: A Decision for a Digital Geogrphic Database", *Proceedings of the Sixth International Symposium on Automated Cartography,* pp. 372–382.
Buchanan, Bruce G. and E. H. Shortlifte, 1985, *Role Based Expert Systems,* Addison–Wesley Publishing Company, Reading, Mass.
Charniak, Eugine et al. 1987, *Artificial Intelligence Porgramming,* Lawrence Erlbaum Associates, Hillsdale, N.J.
Clarke, Keith C. 1986, "Advance in Geographic Information Systems", *Compu. Environ. Urban Systems,* vol. 10, No. 3/4, pp. 175–184.
Cowen, David J. 1988, "GIS vs. CAS vs. DBMS: What are the Differences?", *Photogrammetric Engineering and Remote Sensing,* vol. 54, No. 11, pp. 1551–1555.
Deuker, Kenneth J. 1987, "Geographic Information Systems and Computer–Aided Mapping", *APA Journal* Summer.
Fiedler, David and Bruce H. Hunter 1986, *UNIX System Administration,* Hayden Book Company, Hasbrouck Heights, N.J., p. 58.

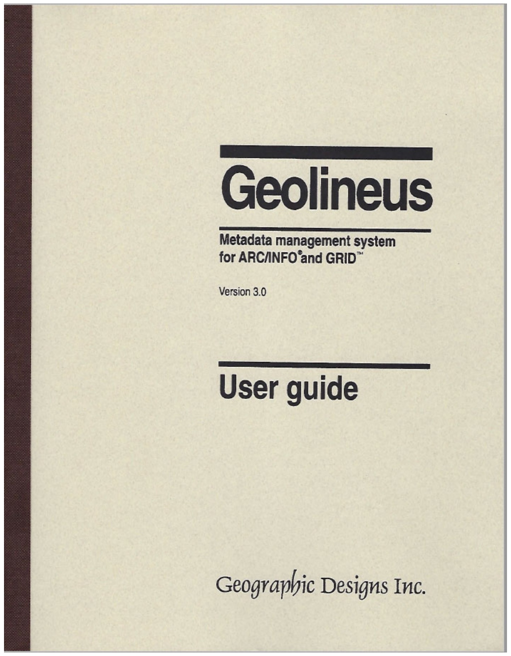
(List continued on next page.)

Primary Examiner—Kevin A. Kriess
Attorney, Agent, or Firm—Jon L. Roberts

[57] **ABSTRACT**

A lineage information processor enables a user to obtain information concerning the various data layers in a spatial data base which contributed to any particular data layer of interest. The component software parses input commands and determines if those commands to the spatial data processing and information systems are valid. The lineage information processor also creates a knowledge representation of the spatial database comprising a meta-database consisting of a semantic network that describes the various data layers in the spatial database and the relationships among these layers. The semantic network consists of parent and child links symbolizing the relationship among data layers, nodes describing the data layers in the spatial database, frames comprising attributes that describe the input data layers, the commands and command modifiers acting on those data layers, and characteristics of the final products. By means of rule-based processing, the lineage information processor does not permit combinations of data layers that are incompatible, and creates commands that can alter incompatible data layers so that the layers can be combined in the desired fashion. A query capability is also provided that enables a user to query in a flexible fashion, the lineage information processor concerning the lineage of data layers in the spatial database.

22 Claims, 68 Drawing Sheets



Geolineus

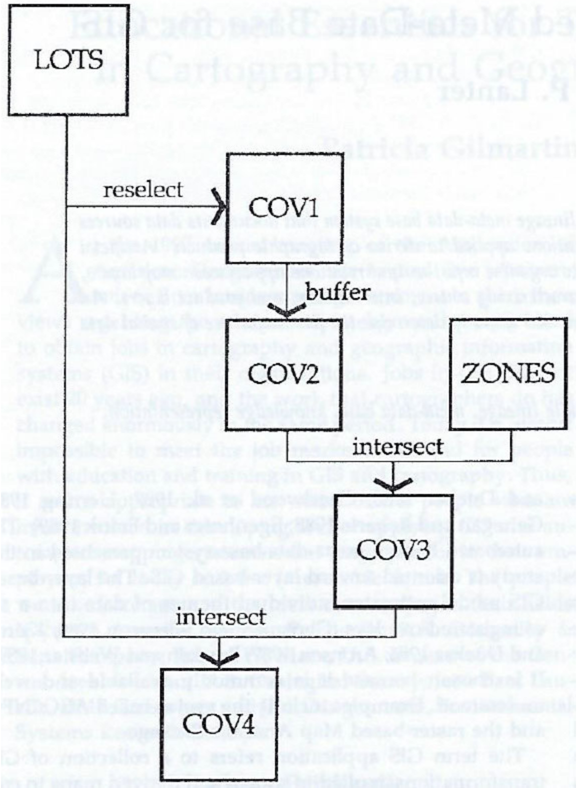
Metadata management system for ARC/INFO® and GRID™

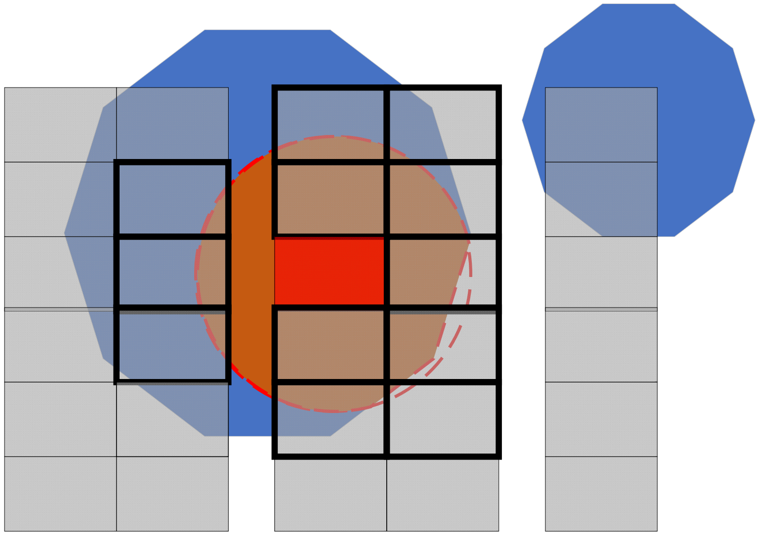
Version 3.0

User guide

Geographic Designs Inc.

Information processing steps in the head of the user as he transformed the LOTS and ZONES datasets to derive COV4…



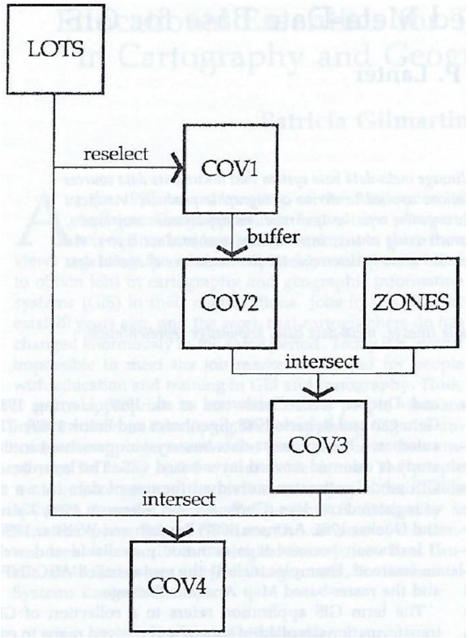


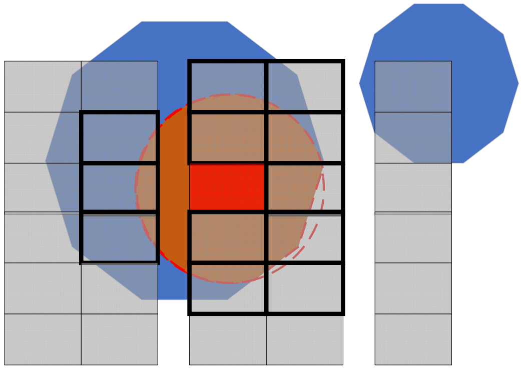
LOTS usually refer to **parcels of land** that are individually owned or designated for specific uses.

ZONES represent **areas with specific land use designations** or characteristics.

A coverage category (COV) in GIS that classifies land based on certain attributes (e.g., vegetation, impervious surfaces, or land cover classification).

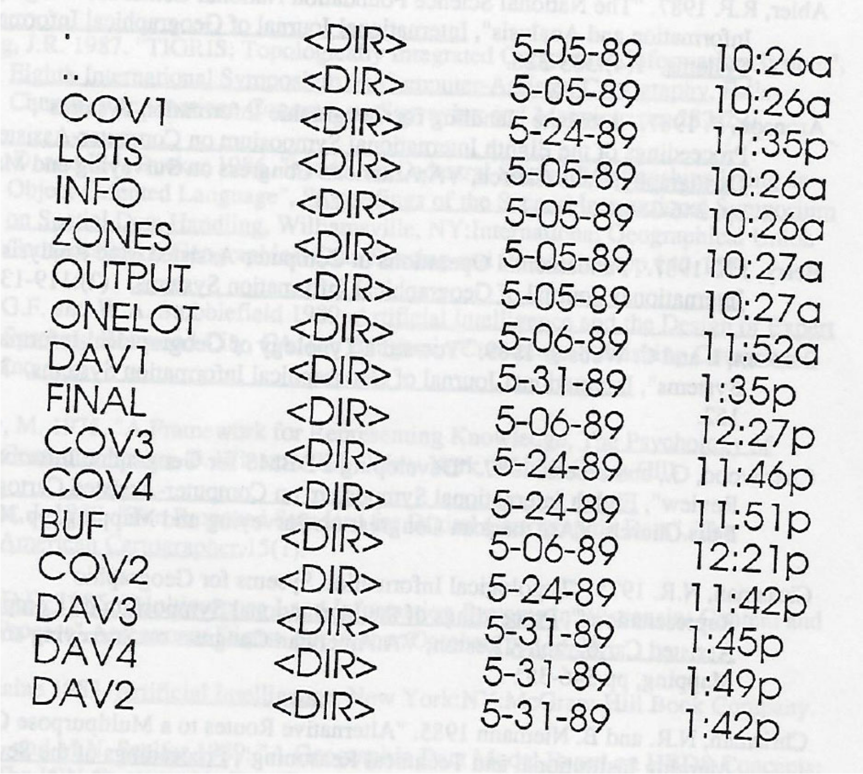
Information processing steps in the head of the user as he transformed the LOTS and ZONES datasets to derive COV4…





Datasets presented by the operating system after data processing concluded…

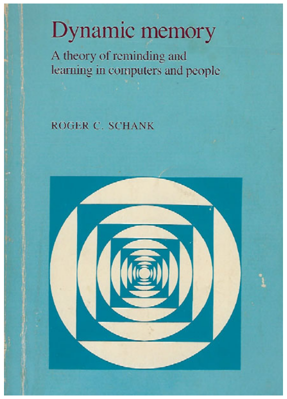
Datasets organized as files in folders

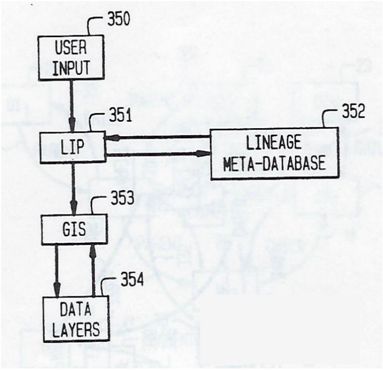


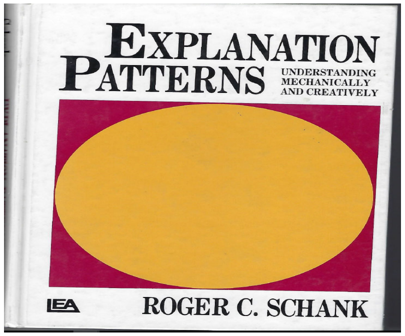
```
.        <DIR>    5-05-89   10:26a
..       <DIR>    5-05-89   10:26a
COV1     <DIR>    5-24-89   11:35p
LOTS     <DIR>    5-05-89   10:26a
INFO     <DIR>    5-05-89   10:26a
ZONES    <DIR>    5-05-89   10:27a
OUTPUT   <DIR>    5-05-89   10:27a
ONELOT   <DIR>    5-06-89   11:52a
DAV1     <DIR>    5-31-89    1:35p
FINAL    <DIR>    5-06-89   12:27p
COV3     <DIR>    5-24-89   11:46p
COV4     <DIR>    5-24-89   11:51p
BUF      <DIR>    5-06-89   12:21p
COV2     <DIR>    5-24-89   11:42p
DAV3     <DIR>    5-31-89    1:45p
DAV4     <DIR>    5-31-89    1:49p
DAV2     <DIR>    5-31-89    1:42p
```

How can I program the computer to help me remember what I knew about the data I was processing when I was processing it?

To help your computer remember what you knew about the data when you were processing it, you need a metadata-driven logging and tracking system. This system should capture context, transformations, decisions, and rationale at each step of data processing.

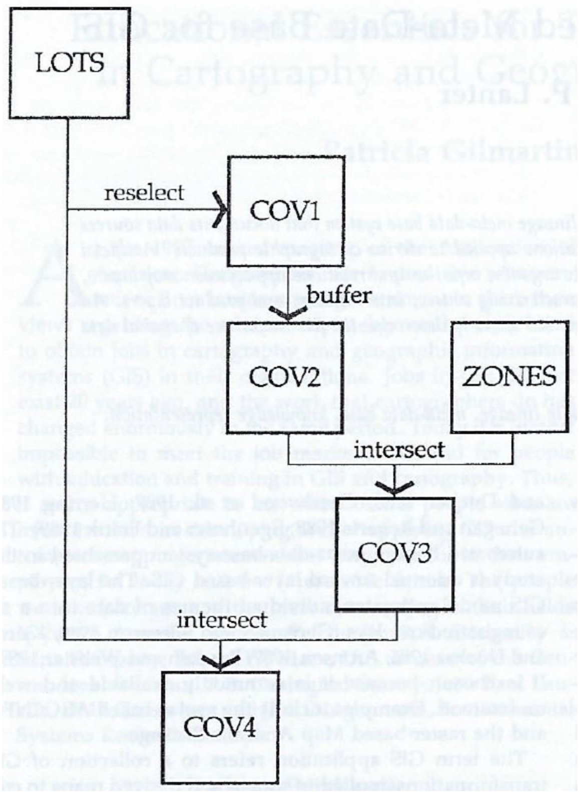


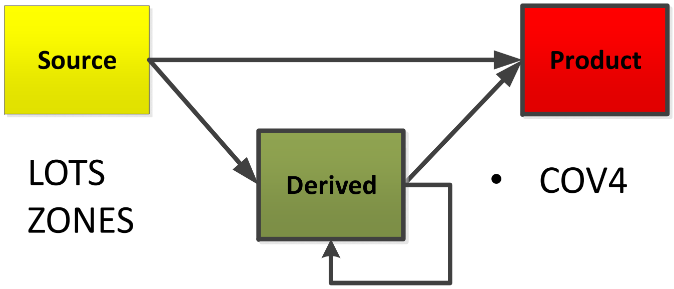




LIP = Lineage Information Processor

How do we understand differences among datasets created during processing applications?

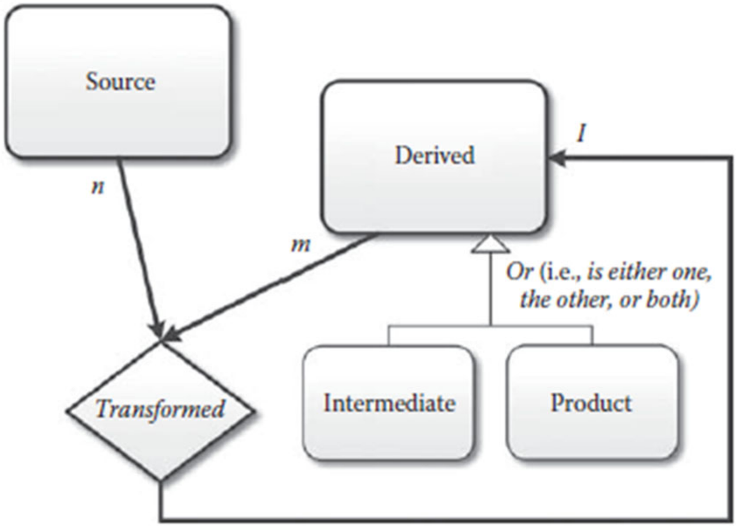




- LOTS
- ZONES

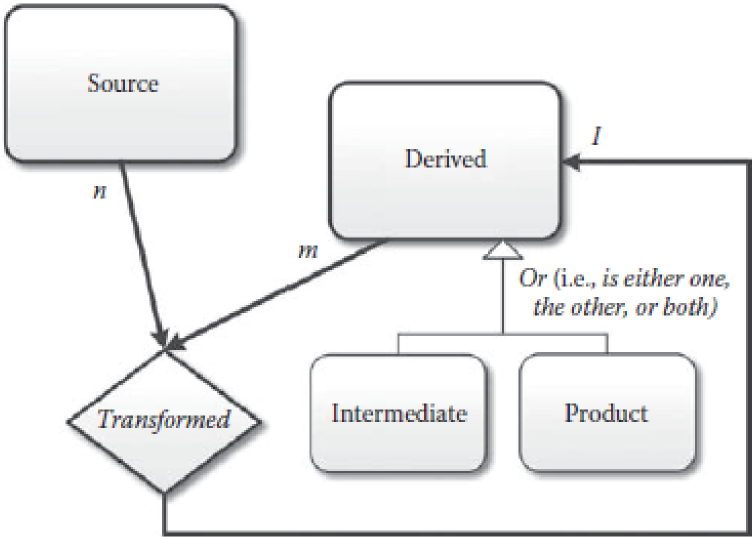
- COV4

- COV1
- COV2
- COV3



Data lineage vocabulary helps communicate how data is processed in an information system

and can aid thinking about how to meet privacy by design requirements

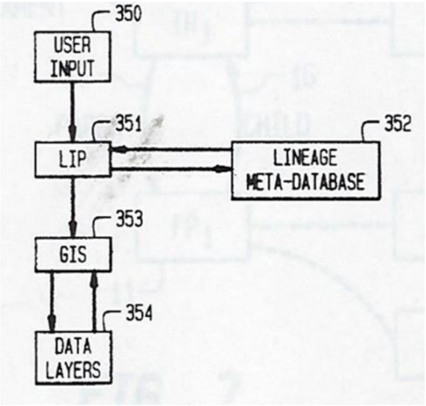


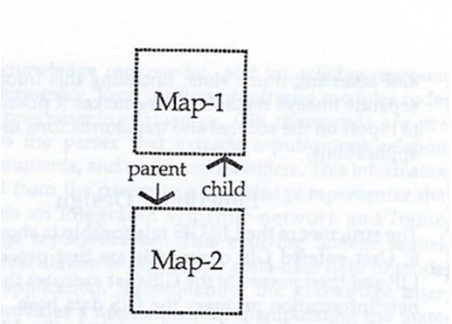
Source datasets *may contain personal data*

Derived datasets inherit this personal data from their input

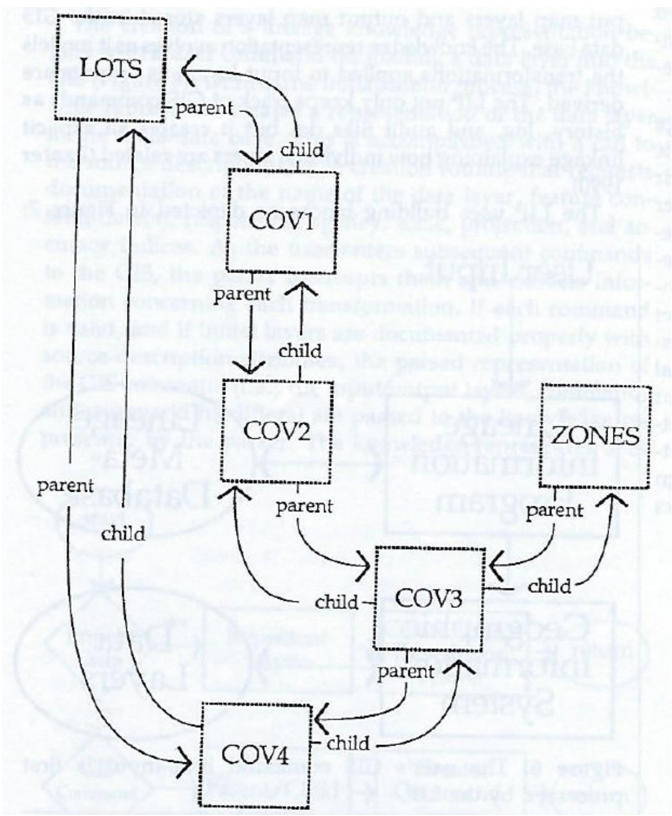
- *Using transformations such as:*
  - *Relational database joins and relates*
  - *Queries, arithmetic, statistical, spatial processing…*

Semantic “parent” & “child” metadata links added to enable deductions about relationships among input & output datasets…





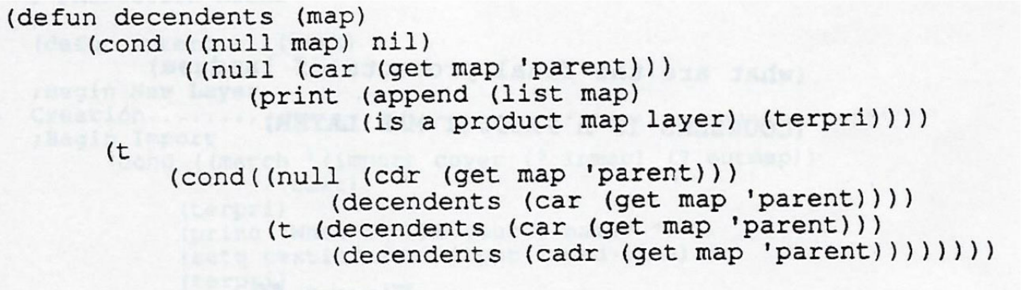
Semantic parent and child links between layers.



Input datasets provided with parent links pointing to output datasets can answer the question: ***Who am I the parent of?***

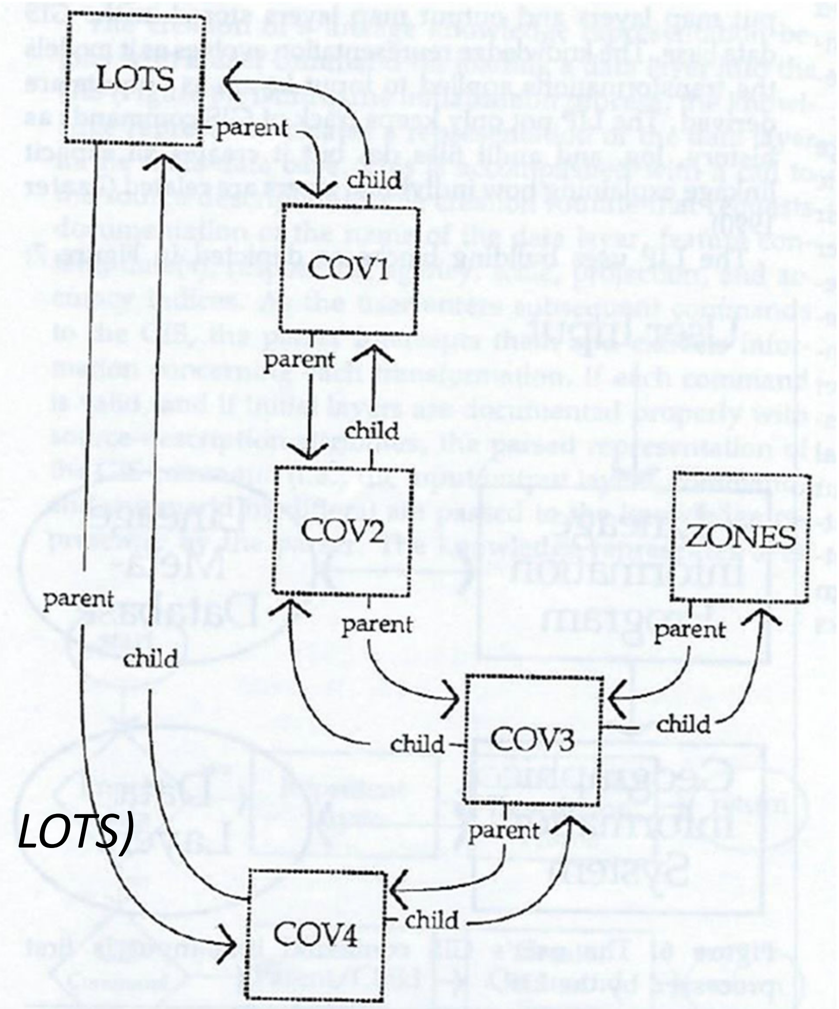
Output datasets’ child links connect them back to their input datasets can answer the question: ***Who am I the child of?***

Descendants function traces parent links to identify all datasets derived from a source or other derived input dataset used within the application.



```
(defun decendents (map)
   (cond ((null map) nil)
         ((null (car (get map 'parent)))
              (print (append (list map)
                       (is a product map layer) (terpri))))
     (t
        (cond((null (cdr (get map 'parent)))
                  (decendents (car (get map 'parent))))
             (t (decendents (car (get map 'parent)))
                  (decendents (cadr (get map 'parent)))))))
```

Descendants (“LOTS”) = (COV1, COV2, COV3, COV4)

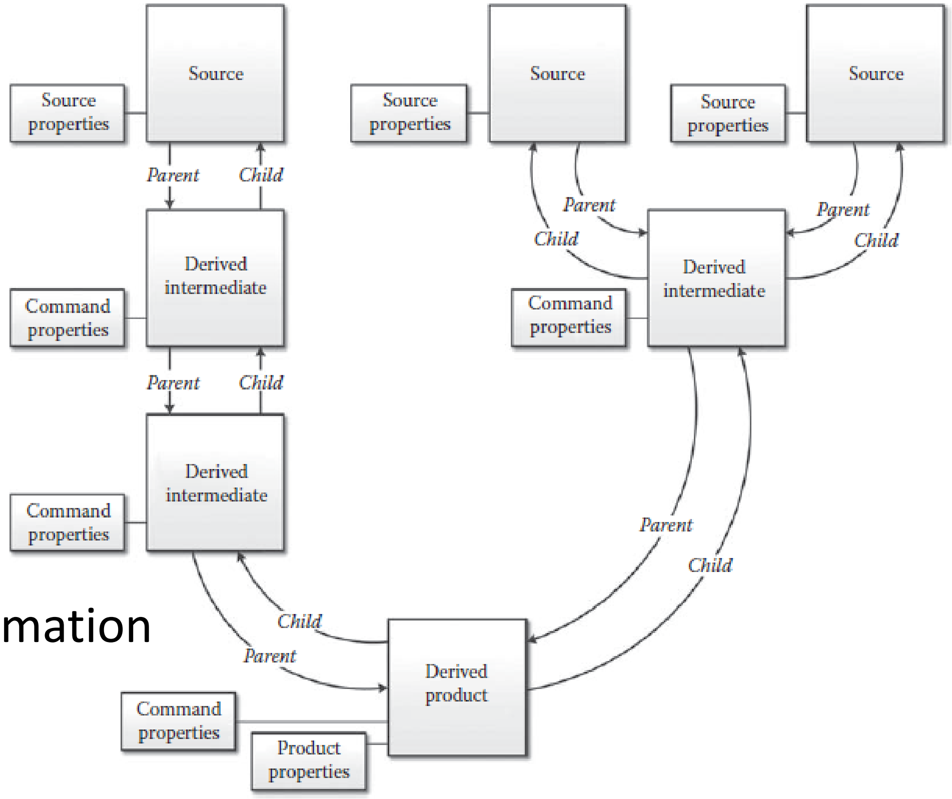


Ancestors function traces child links to identify input datasets used to create a derived dataset

Ancestors (“COV4”) = (LOTS, COV3, ZONES, COV2, COV1, LOTS)

Source properties can include:

- Originating organization
- Data content (i.e. entity and attribute definitions)
- Timeliness (e.g. when collected, when acquired,…)
- Accuracy
- Confidentiality security categorization of attributes
  - Privacy sensitivity of attributes
- Integrity categorization of attributes…
- Availability categorization…

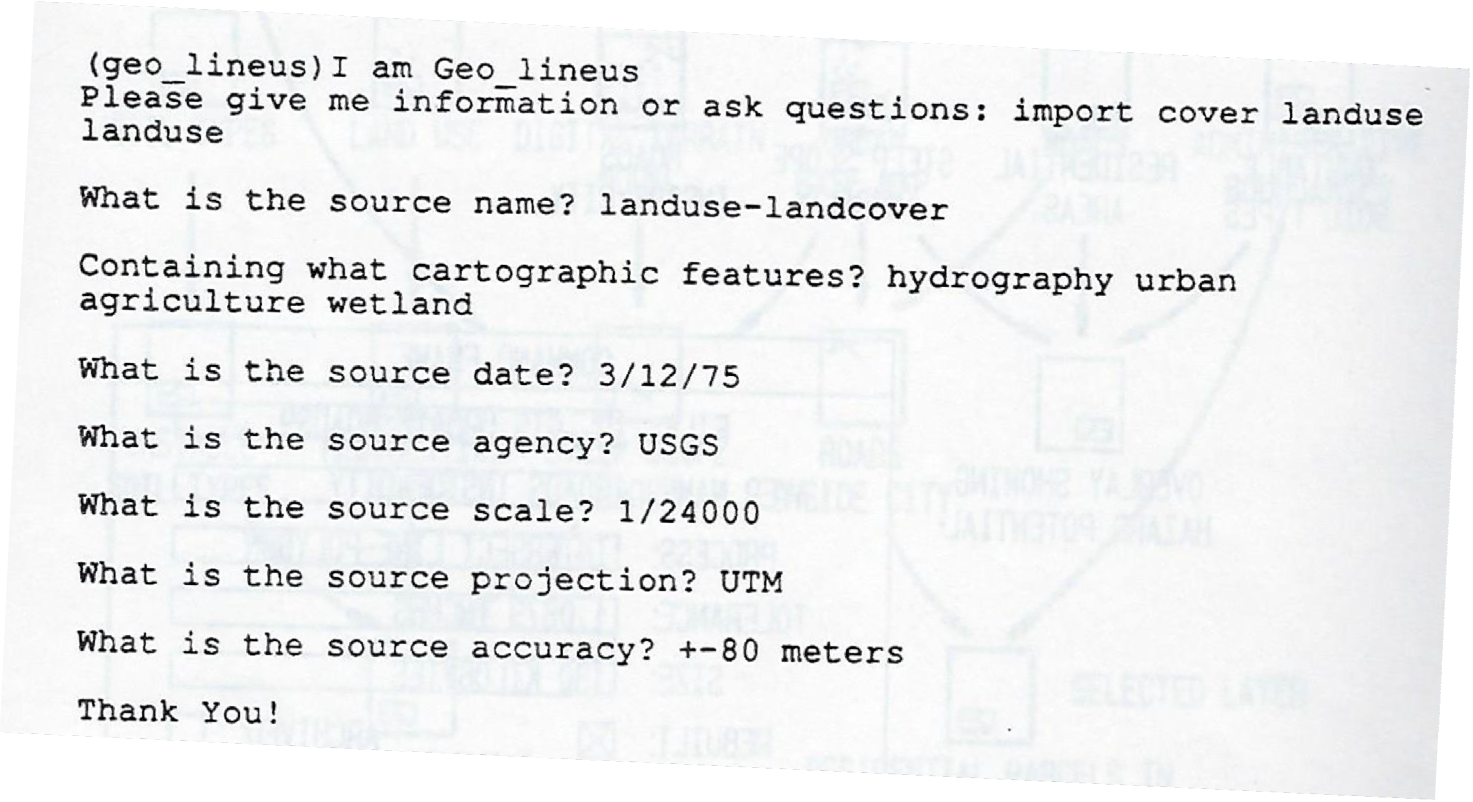


Command properties include details of the transformation

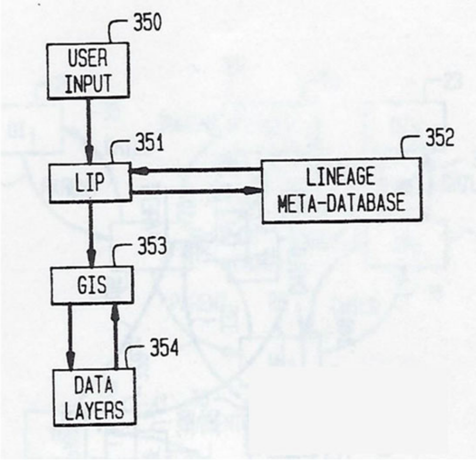
Product properties include the product's

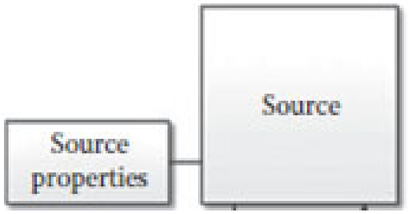
- intended goal
- Users
- when published
- responsible manager,…

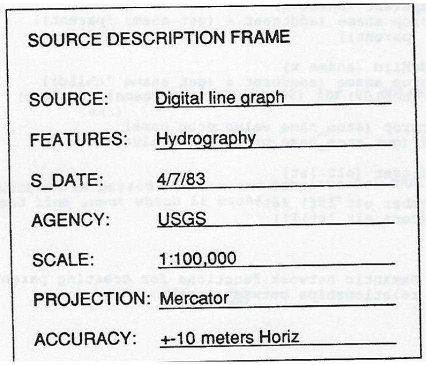
Meet Geo_lineus
source metadata input



```
(geo_lineus)I am Geo_lineus
Please give me information or ask questions: import cover landuse
landuse

What is the source name? landuse-landcover

Containing what cartographic features? hydrography urban
agriculture wetland

What is the source date? 3/12/75

What is the source agency? USGS

What is the source scale? 1/24000

What is the source projection? UTM

What is the source accuracy? +-80 meters

Thank You!
```







SOURCE DESCRIPTION FRAME

SOURCE: Digital line graph

FEATURES: Hydrography

S_DATE: 4/7/83

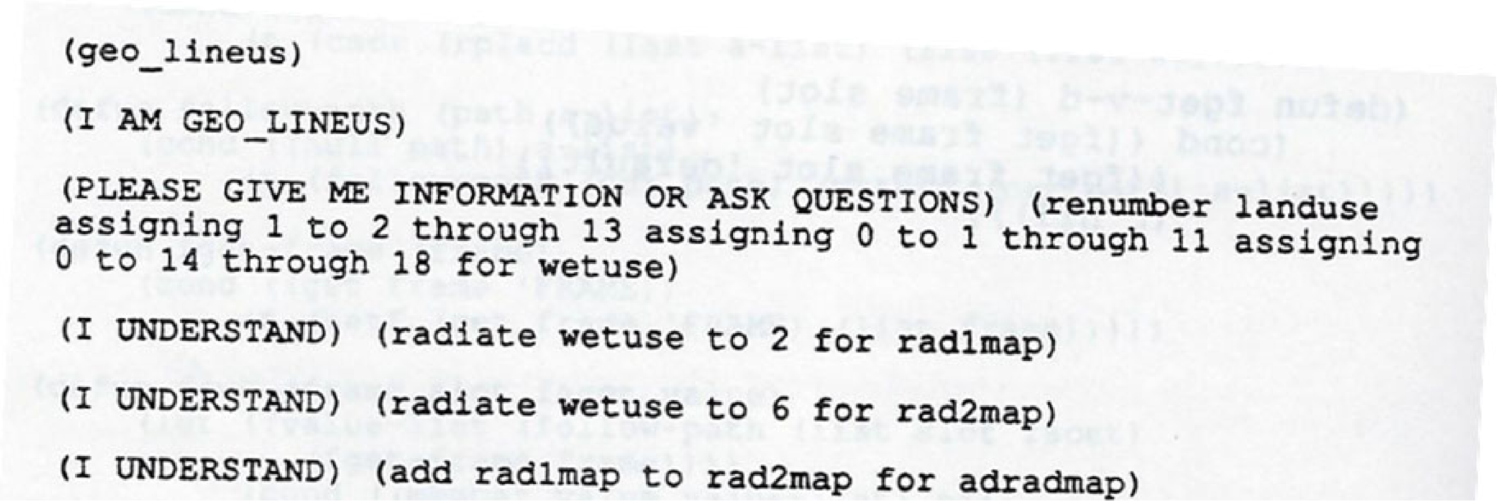
AGENCY: USGS

SCALE: 1:100,000

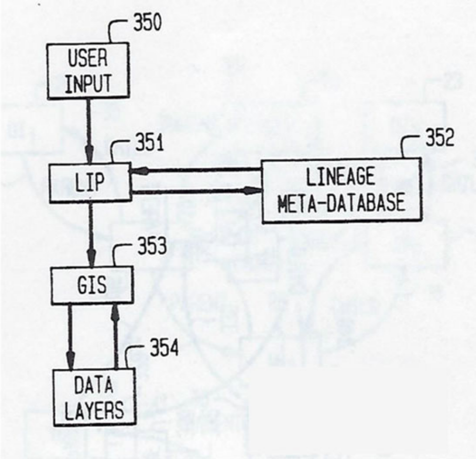
PROJECTION: Mercator

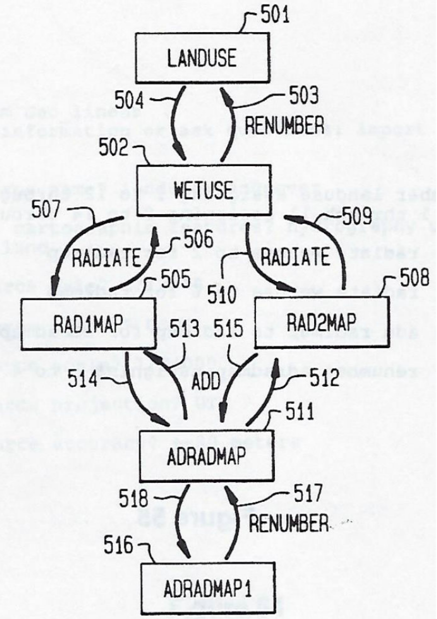
ACCURACY: +-10 meters Horiz

Command metadata input…

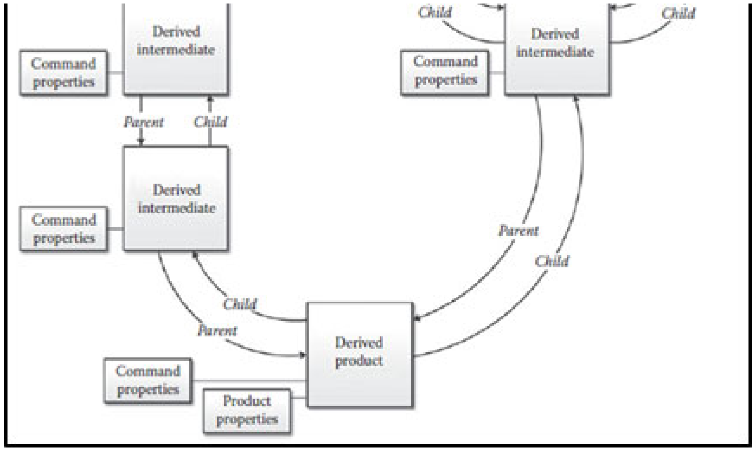


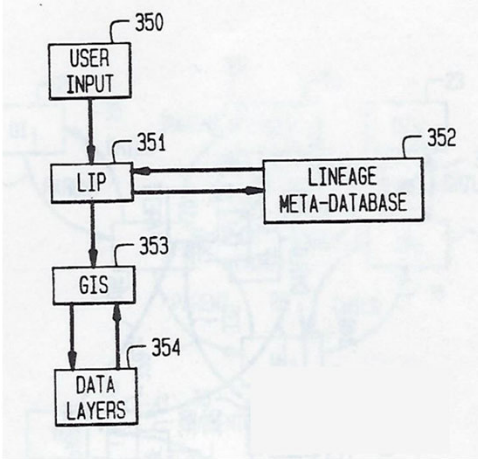
```
(geo_lineus)

(I AM GEO_LINEUS)

(PLEASE GIVE ME INFORMATION OR ASK QUESTIONS) (renumber landuse
assigning 1 to 2 through 13 assigning 0 to 1 through 11 assigning
0 to 14 through 18 for wetuse)

(I UNDERSTAND) (radiate wetuse to 2 for rad1map)

(I UNDERSTAND) (radiate wetuse to 6 for rad2map)

(I UNDERSTAND) (add rad1map to rad2map for adradmap)
```

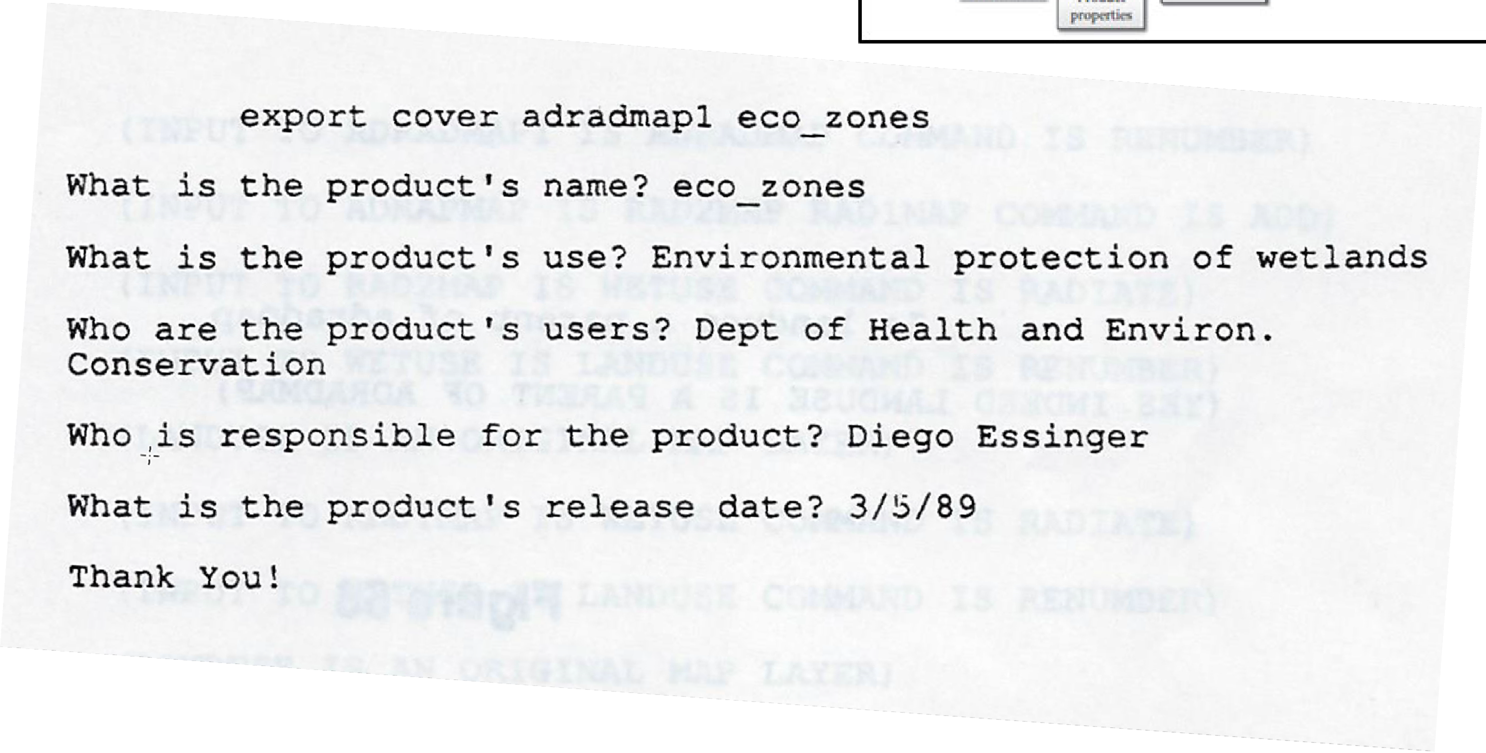




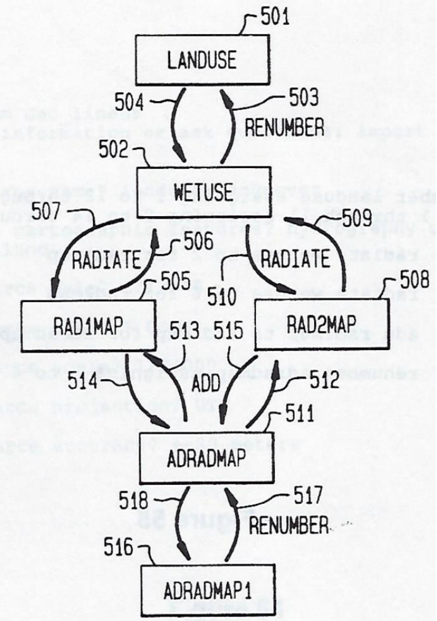
Product Metadata input…



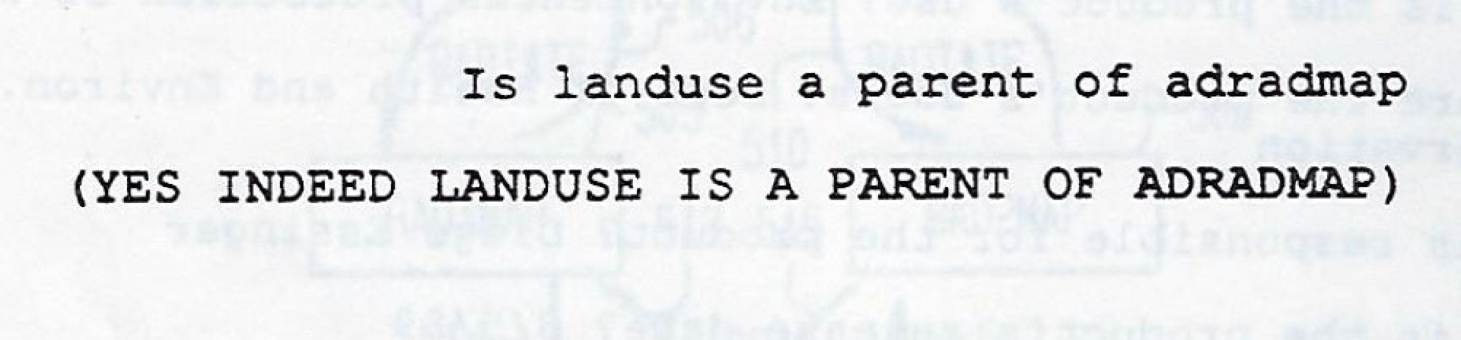




```
      export cover adradmap1 eco_zones

What is the product's name? eco_zones

What is the product's use? Environmental protection of wetlands

Who are the product's users? Dept of Health and Environ.
Conservation

Who is responsible for the product? Diego Essinger

What is the product's release date? 3/5/89

Thank You!
```

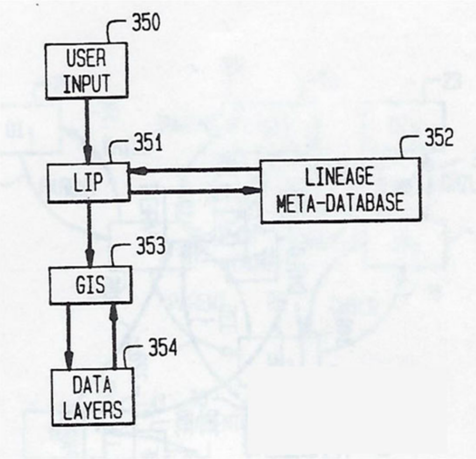


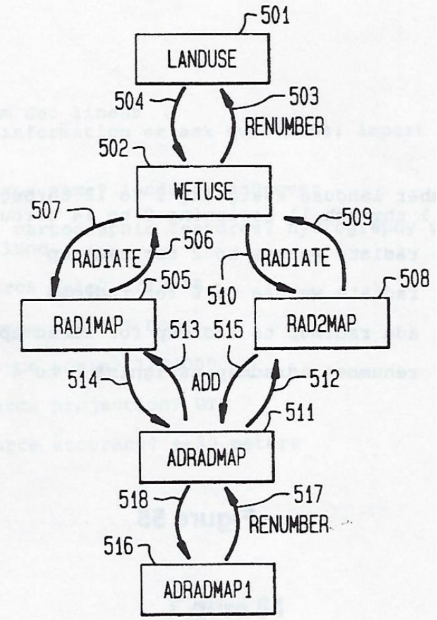
Querying metadata…



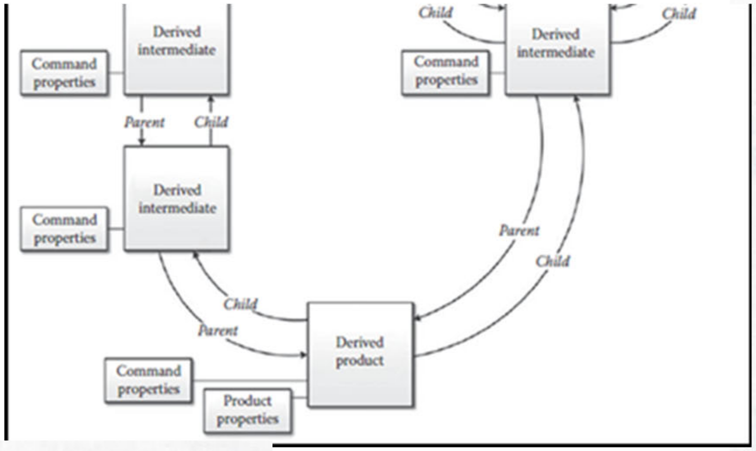
Is landuse a parent of adradmap

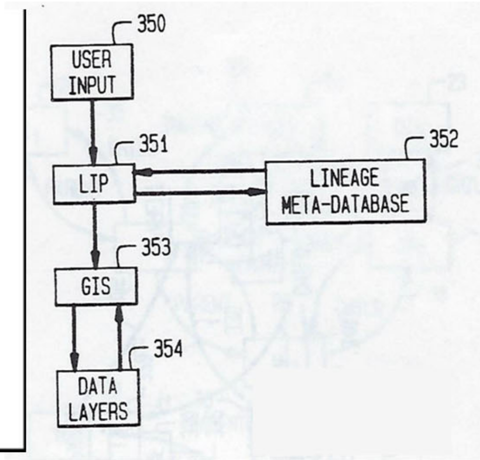
(YES INDEED LANDUSE IS A PARENT OF ADRADMAP)

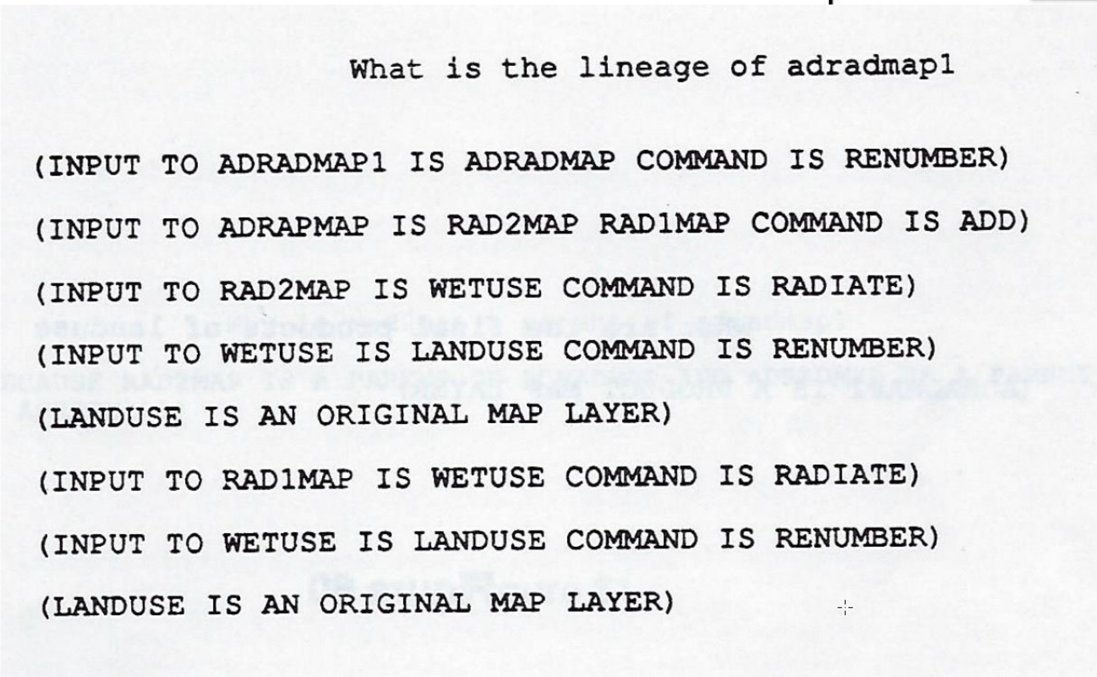




Querying metadata…







What is the lineage of adradmap1

(INPUT TO ADRADMAP1 IS ADRADMAP COMMAND IS RENUMBER)

(INPUT TO ADRAPMAP IS RAD2MAP RAD1MAP COMMAND IS ADD)

(INPUT TO RAD2MAP IS WETUSE COMMAND IS RADIATE)

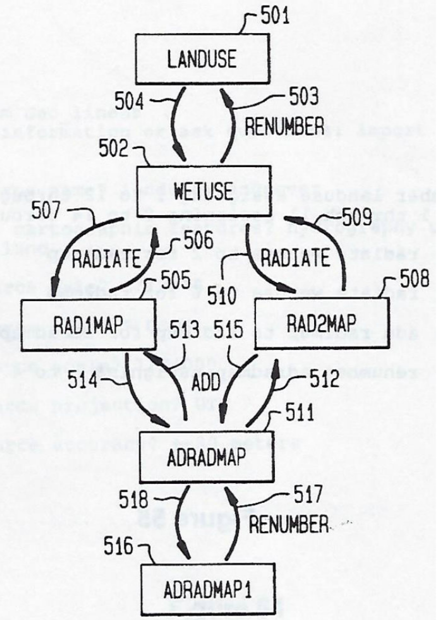
(INPUT TO WETUSE IS LANDUSE COMMAND IS RENUMBER)

(LANDUSE IS AN ORIGINAL MAP LAYER)

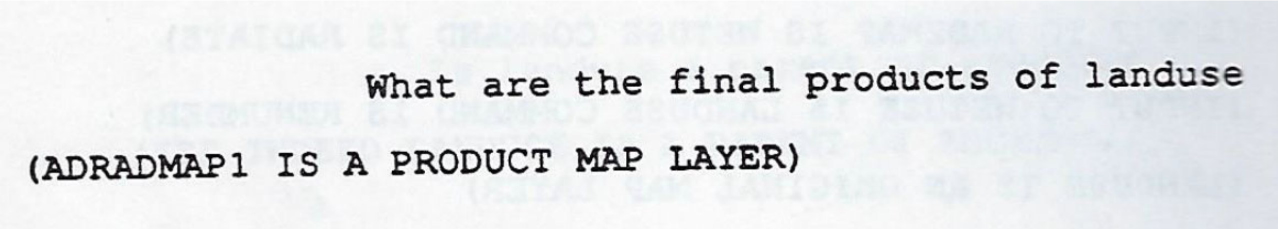
(INPUT TO RAD1MAP IS WETUSE COMMAND IS RADIATE)

(INPUT TO WETUSE IS LANDUSE COMMAND IS RENUMBER)

(LANDUSE IS AN ORIGINAL MAP LAYER)

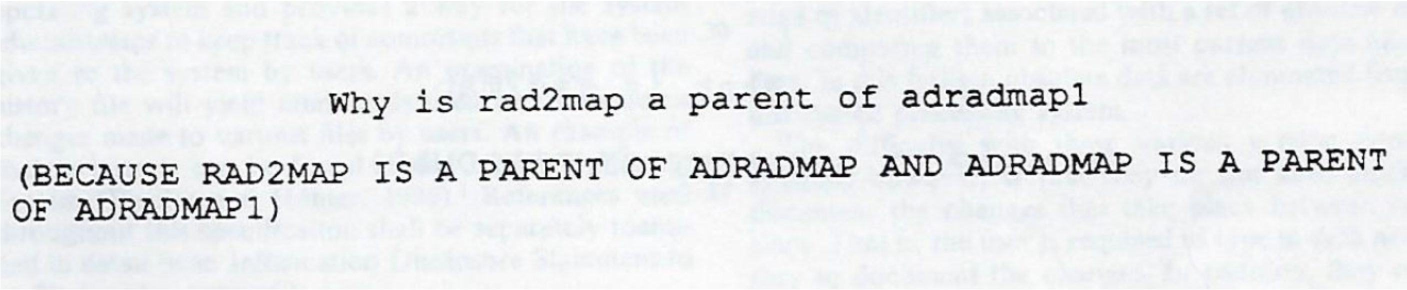


Querying metadata…



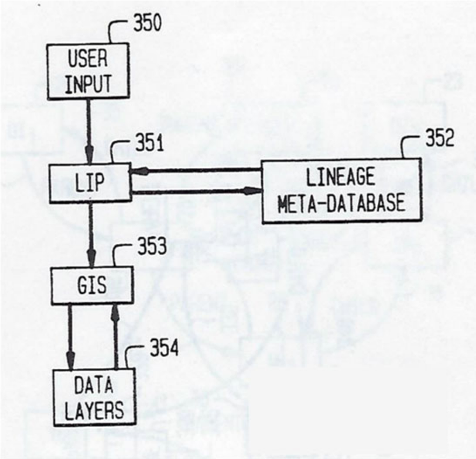
What are the final products of landuse

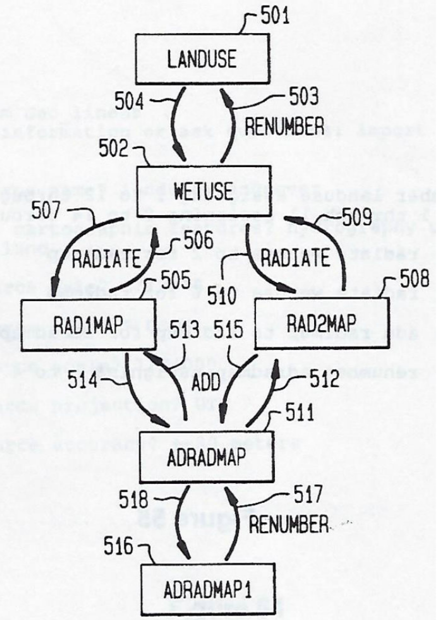
(ADRADMAP1 IS A PRODUCT MAP LAYER)



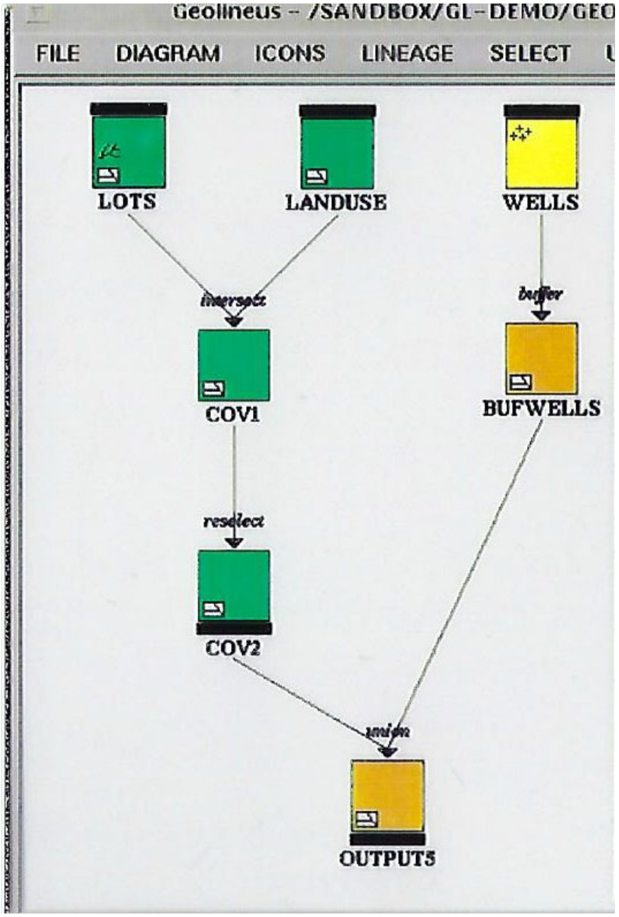
Why is rad2map a parent of adradmap1

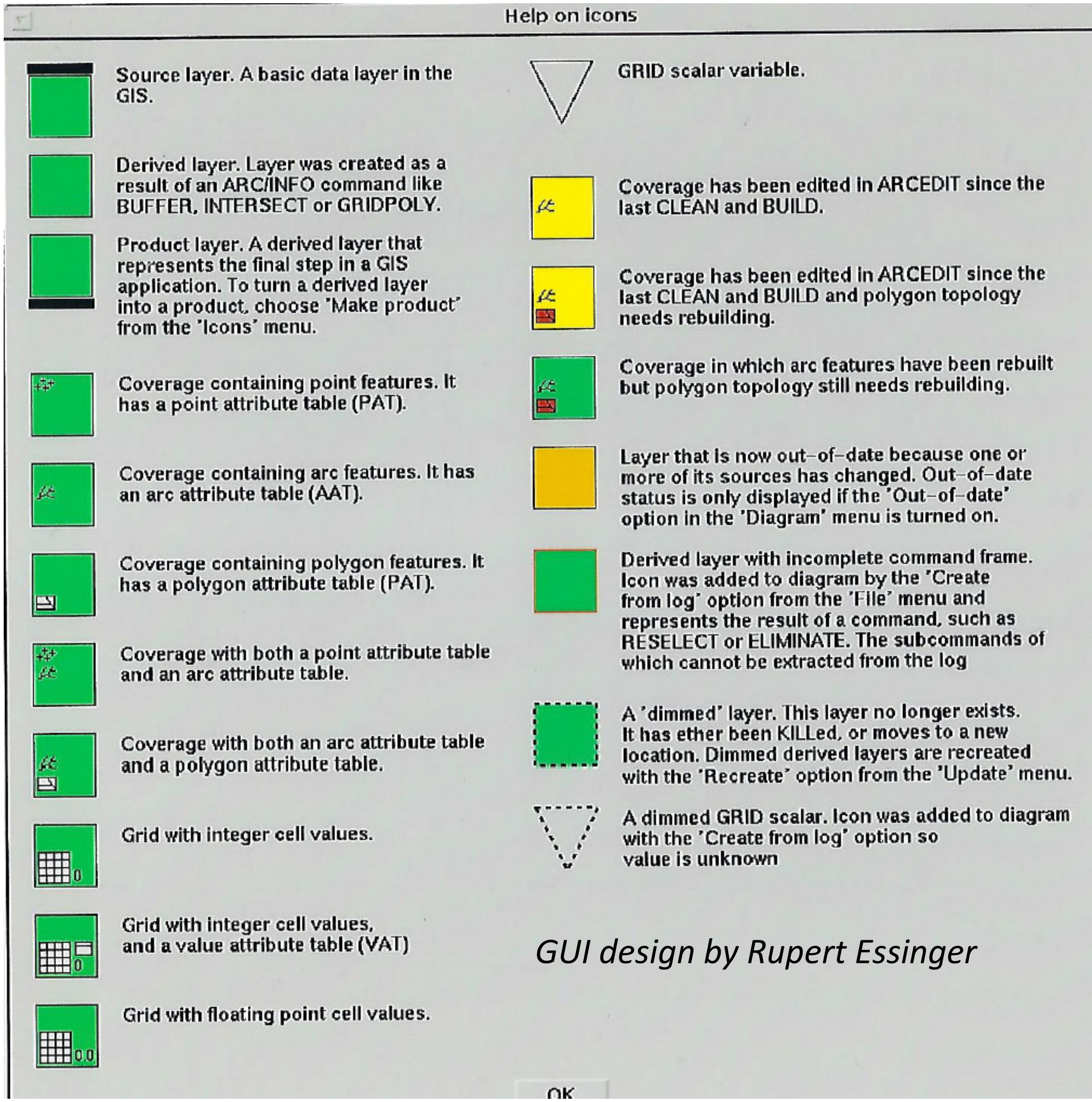
(BECAUSE RAD2MAP IS A PARENT OF ADRADMAP AND ADRADMAP IS A PARENT OF ADRADMAP1)





Adding a graphical user interface…




Help on icons

🟩 Source layer. A basic data layer in the GIS.

🟩 Derived layer. Layer was created as a result of an ARC/INFO command like BUFFER, INTERSECT or GRIDPOLY.

🟩 Product layer. A derived layer that represents the final step in a GIS application. To turn a derived layer into a product, choose 'Make product' from the 'Icons' menu.

🟩 Coverage containing point features. It has a point attribute table (PAT).

🟩 Coverage containing arc features. It has an arc attribute table (AAT).

🟩 Coverage containing polygon features. It has a polygon attribute table (PAT).

🟩 Coverage with both a point attribute table and an arc attribute table.

🟩 Coverage with both an arc attribute table and a polygon attribute table.

🟩 Grid with integer cell values.

🟩 Grid with integer cell values, and a value attribute table (VAT)

🟩 Grid with floating point cell values.

▽ GRID scalar variable.

🟨 Coverage has been edited in ARCEDIT since the last CLEAN and BUILD.

🟨 Coverage has been edited in ARCEDIT since the last CLEAN and BUILD and polygon topology needs rebuilding.

🟩 Coverage in which arc features have been rebuilt but polygon topology still needs rebuilding.

🟧 Layer that is now out-of-date because one or more of its sources has changed. Out-of-date status is only displayed if the 'Out-of-date' option in the 'Diagram' menu is turned on.

🟩 Derived layer with incomplete command frame. Icon was added to diagram by the 'Create from log' option from the 'File' menu and represents the result of a command, such as RESELECT or ELIMINATE. The subcommands of which cannot be extracted from the log

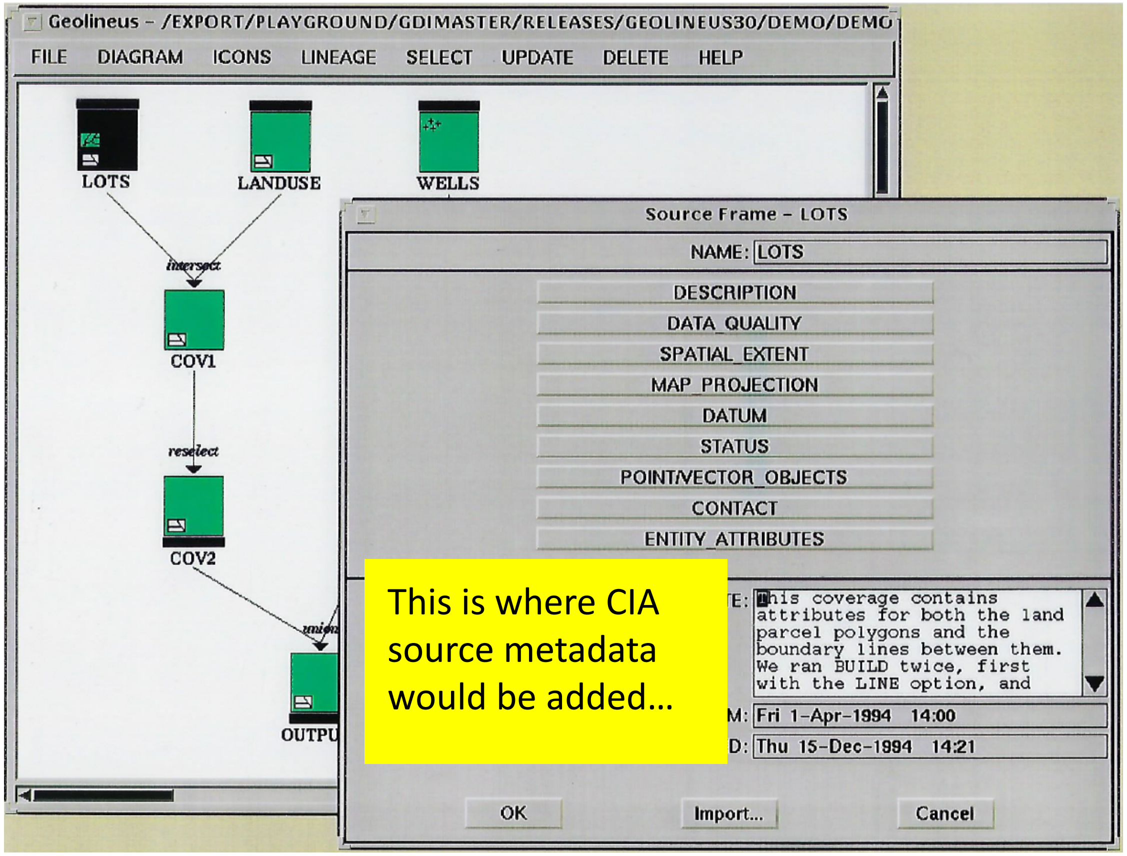
🟩 A 'dimmed' layer. This layer no longer exists. It has ether been KILLed, or moves to a new location. Dimmed derived layers are recreated with the 'Recreate' option from the 'Update' menu.

▽ A dimmed GRID scalar. Icon was added to diagram with the 'Create from log' option so value is unknown

OK

GUI design by Rupert Essinger

Working with source and command metadata



Source Frame – LOTS

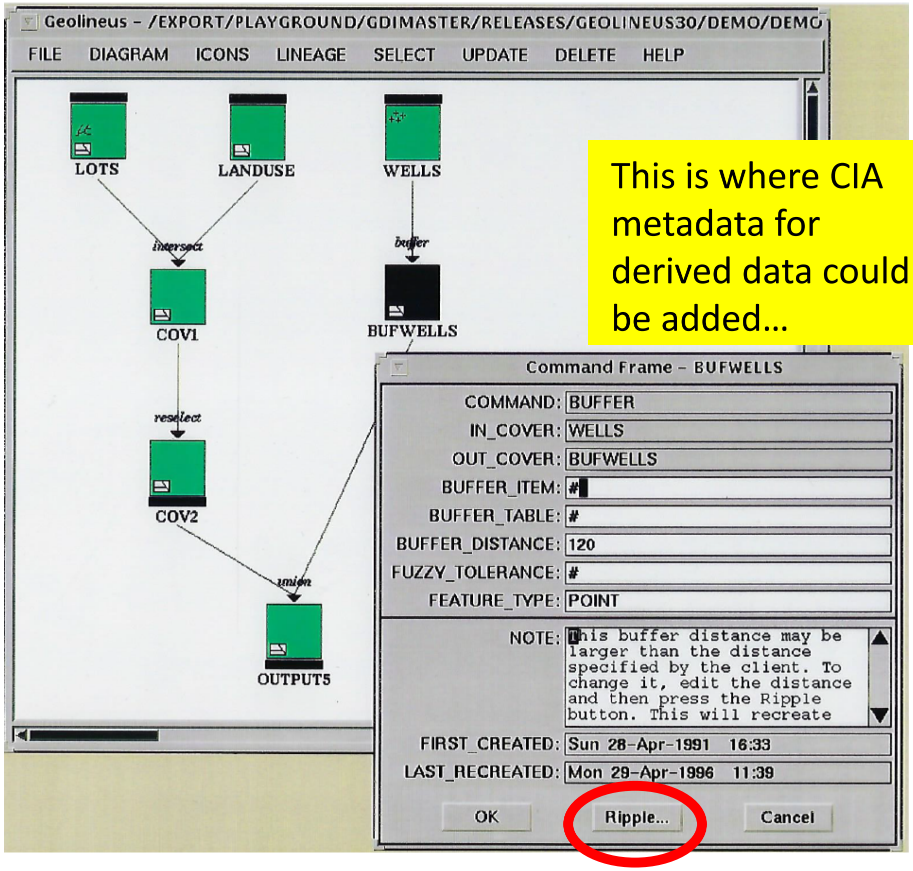
NAME: LOTS

TE: This coverage contains attributes for both the land parcel polygons and the boundary lines between them. We ran BUILD twice, first with the LINE option, and

M: Fri 1-Apr-1994 14:00

D: Thu 15-Dec-1994 14:21

This is where CIA source metadata would be added…



Command Frame – BUFWELLS

COMMAND: BUFFER

IN_COVER: WELLS

OUT_COVER: BUFWELLS

BUFFER_ITEM: #

BUFFER_TABLE: #

BUFFER_DISTANCE: 120

FUZZY_TOLERANCE: #

FEATURE_TYPE: POINT

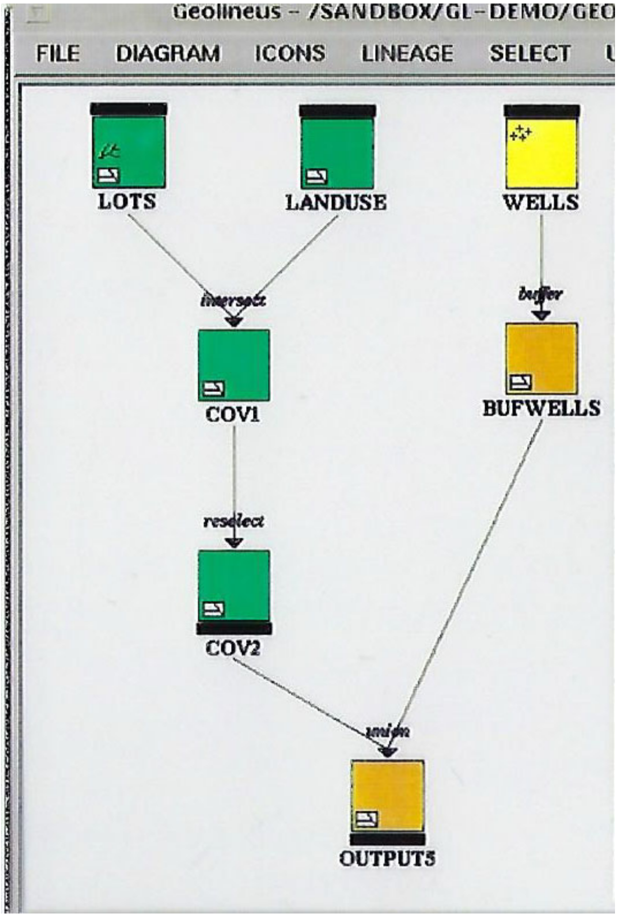
NOTE: This buffer distance may be larger than the distance specified by the client. To change it, edit the distance and then press the Ripple button. This will recreate

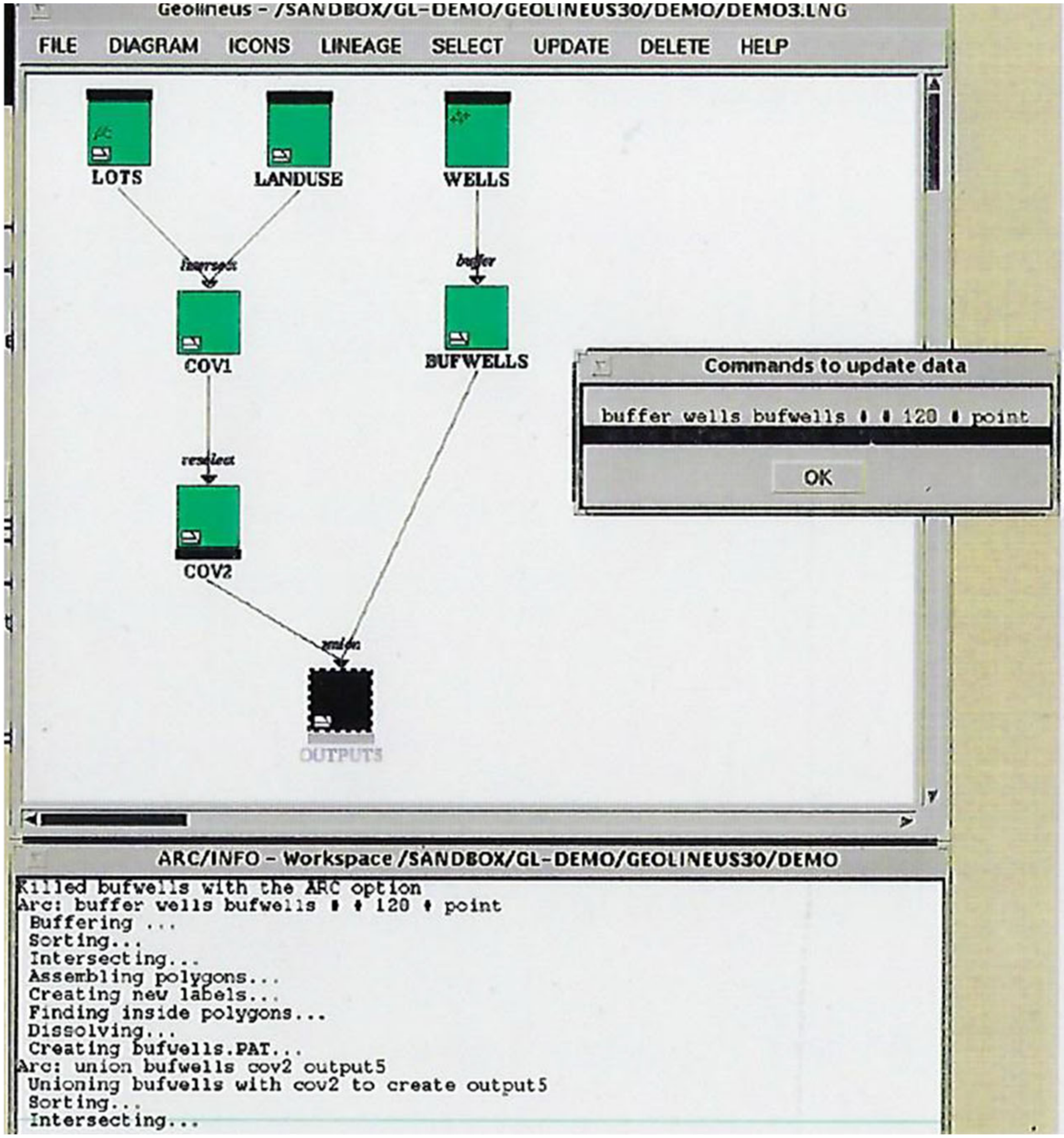
FIRST_CREATED: Sun 28-Apr-1991 16:33

LAST_RECREATED: Mon 29-Apr-1996 11:39

This is where CIA metadata for derived data could be added…

Update propagation…





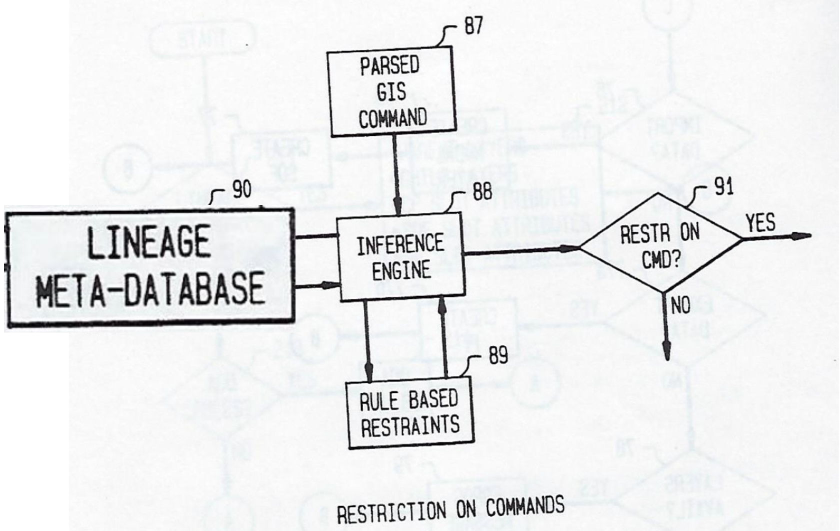
Commands to update data

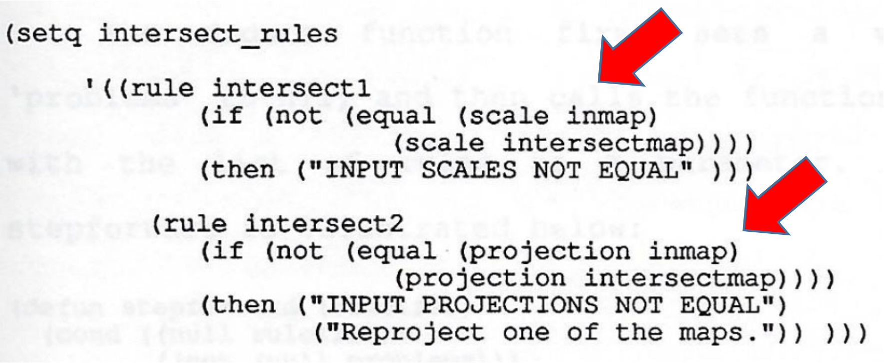
```
buffer wells bufwells # # 120 # point
```

ARC/INFO - Workspace /SANDBOX/GL-DEMO/GEOLINEUS30/DEMO

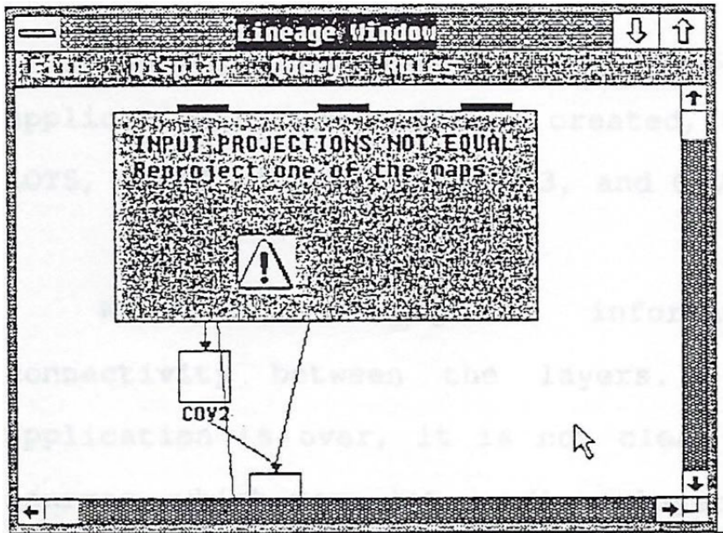
```
Killed bufwells with the ARC option
Arc: buffer wells bufwells # # 120 # point
 Buffering ...
 Sorting...
 Intersecting...
 Assembling polygons...
 Creating new labels...
 Finding inside polygons...
 Dissolving...
 Creating bufwells.PAT...
Arc: union bufwells cov2 output5
 Unioning bufwells with cov2 to create output5
 Sorting...
 Intersecting...
```

Data source metadata based integrity constraint

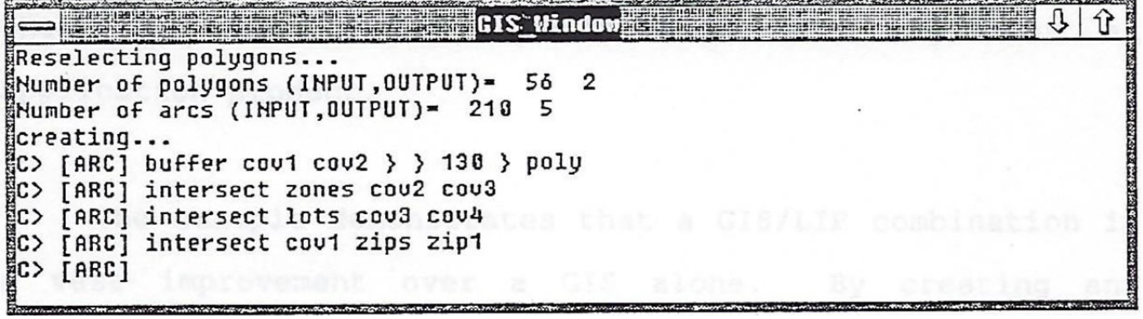




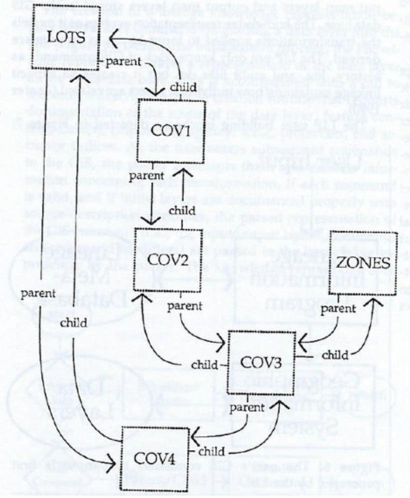
```
(setq intersect_rules
    '((rule intersect1
          (if (not (equal (scale inmap)
                          (scale intersectmap))))
          (then ("INPUT SCALES NOT EQUAL" )))

      (rule intersect2
          (if (not (equal (projection inmap)
                          (projection intersectmap))))
          (then ("INPUT PROJECTIONS NOT EQUAL")
                ("Reproject one of the maps.")) )))
```



INPUT PROJECTIONS NOT EQUAL.
Reproject one of the maps.

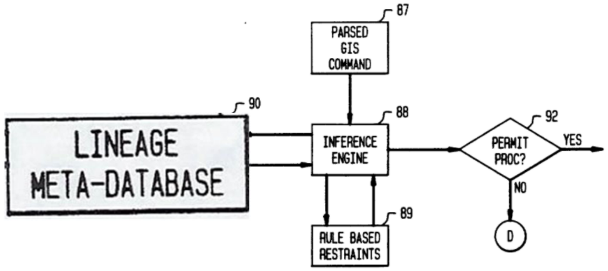


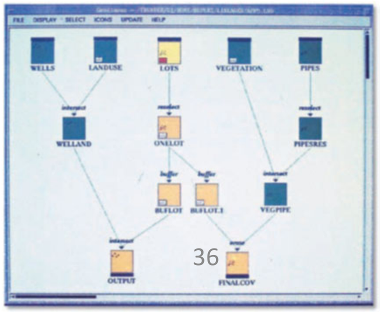
```
Reselecting polygons...
Number of polygons (INPUT,OUTPUT)=  56  2
Number of arcs (INPUT,OUTPUT)=  210  5
creating...
C> [ARC] buffer cov1 cov2 } } 130 } poly
C> [ARC] intersect zones cov2 cov3
C> [ARC] intersect lots cov3 cov4
C> [ARC] intersect cov1 zips zip1
C> [ARC]
```



Data lineage metadata can help information systems meet key data privacy by design requirements, including:

- Enabling data subjects access, review and rectify their personal data?
- Enable data subjects to withdraw given consent with effect for the future by:
    a. Blocking access to their personal data?
    b. Constraining processing and usage of their personal data?
    c. Erasing their personal data?
- Blocking and restricting personal data obtained for one purpose from being processed for other purposes not compatible with the original purpose



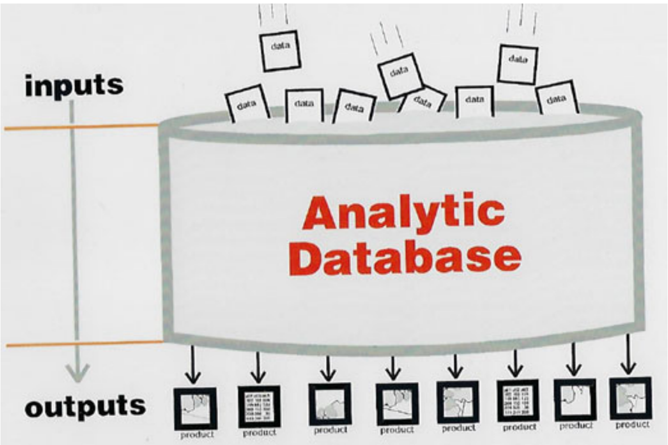


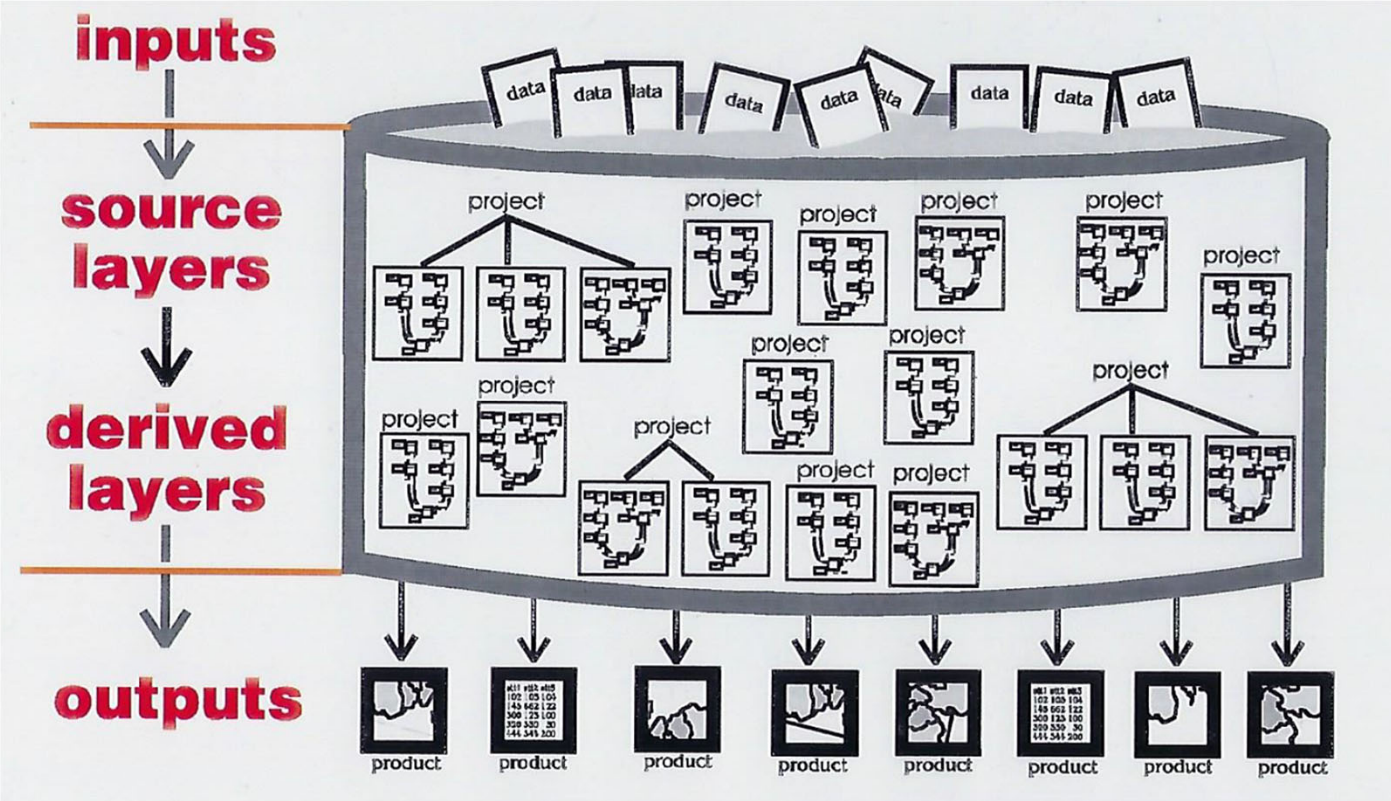
Case Study: Data lineage metadata enabled audit at Southern California Edison

Focus of the audit:

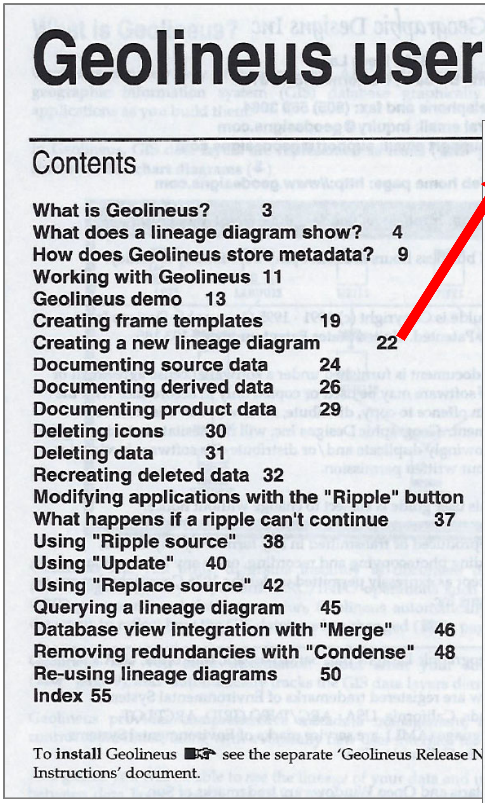
1. Documentation and understandingof GIS decision support data
2. Replicability of data used in decision making

Data provenance audit problem…





Metadata Analysis of data and processing



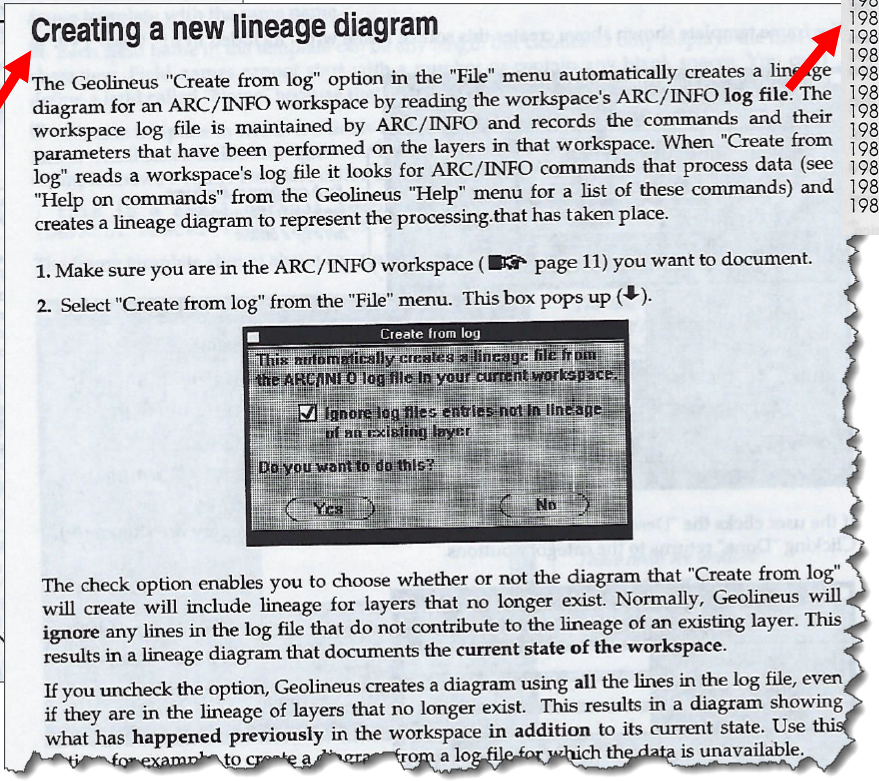
Geolineus user guide

Contents

What is Geolineus? 3
What does a lineage diagram show? 4
How does Geolineus store metadata? 9
Working with Geolineus 11
Geolineus demo 13
Creating frame templates 19
Creating a new lineage diagram 22
Documenting source data 24
Documenting derived data 26
Documenting product data 29
Deleting icons 30
Deleting data 31
Recreating deleted data 32
Modifying applications with the "Ripple" button
What happens if a ripple can't continue 37
Using "Ripple source" 38
Using "Update" 40
Using "Replace source" 42
Querying a lineage diagram 45
Database view integration with "Merge" 46
Removing redundancies with "Condense" 48
Re-using lineage diagrams 50
Index 55

To **install** Geolineus ☞ see the separate 'Geolineus Release N

Instructions' document.



Creating a new lineage diagram

The Geolineus "Create from log" option in the "File" menu automatically creates a lineage diagram for an ARC/INFO workspace by reading the workspace's ARC/INFO **log file**. The workspace log file is maintained by ARC/INFO and records the commands and their parameters that have been performed on the layers in that workspace. When "Create from log" reads a workspace's log file it looks for ARC/INFO commands that process data (see "Help on commands" from the Geolineus "Help" menu for a list of these commands) and creates a lineage diagram to represent the processing.that has taken place.

1. Make sure you are in the ARC/INFO workspace (☞ page 11) you want to document.

2. Select "Create from log" from the "File" menu. This box pops up (⬇).

Create from log

This automatically creates a lineage file from the ARC/INFO log file in your current workspace.

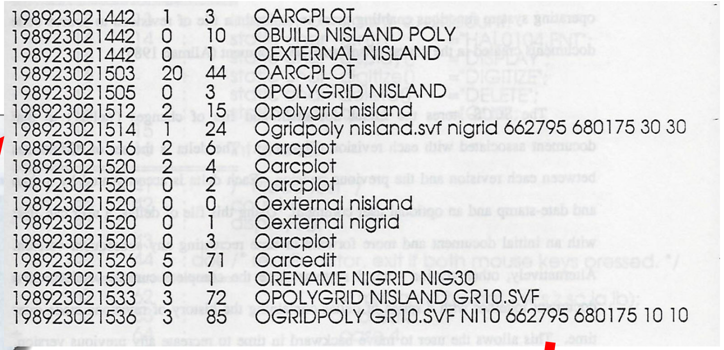
☑ Ignore log files entries not in lineage of an existing layer

Do you want to do this?

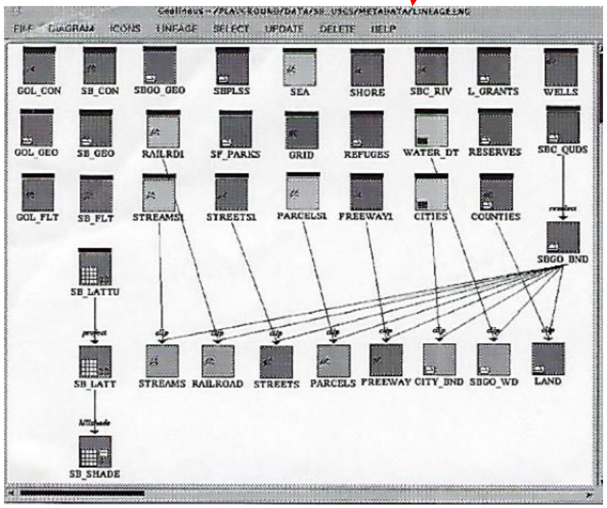
The check option enables you to choose whether or not the diagram that "Create from log" will create will include lineage for layers that no longer exist. Normally, Geolineus will **ignore** any lines in the log file that do not contribute to the lineage of an existing layer. This results in a lineage diagram that documents the **current state of the workspace**.

If you uncheck the option, Geolineus creates a diagram using **all** the lines in the log file, even if they are in the lineage of layers that no longer exist. This results in a diagram showing what has **happened previously** in the workspace **in addition** to its current state. Use this option, for example, to create a diagram from a log file for which the data is unavailable.

Log Files



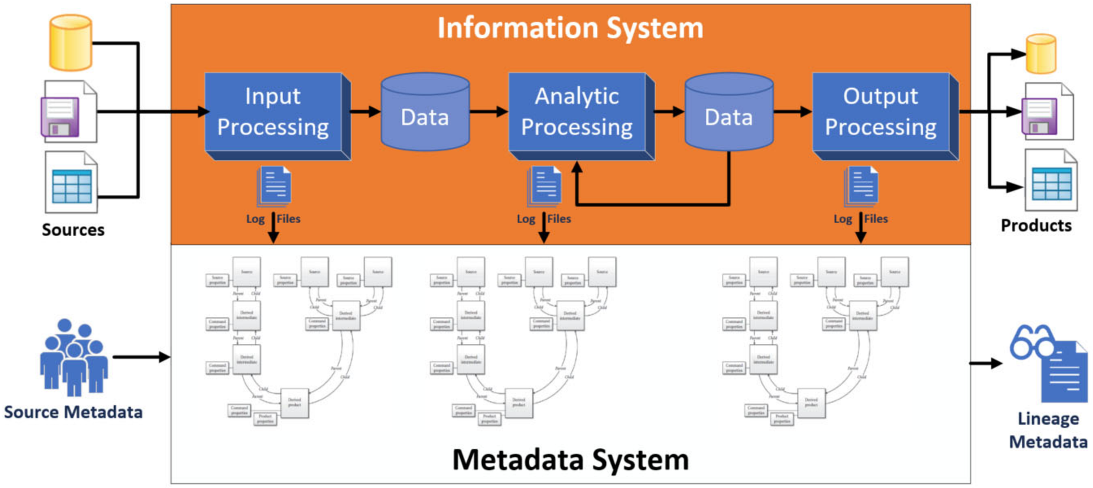
```
198923021442   1    3    OARCPLOT
198923021442   0    10   OBUILD NISLAND POLY
198923021442   0    1    OEXTERNAL NISLAND
198923021503   20   44   OARCPLOT
198923021505   0    3    OPOLYGRID NISLAND
198923021512   2    15   Opolygrid nisland
198923021514   1    24   Ogridpoly nisland.svf nigrid 662795 680175 30 30
198923021516   2    6    Oarcplot
198923021520   2    4    Oarcplot
198923021520   0    2    Oarcplot
198923021520   0    0    Oexternal nisland
198923021520   0    1    Oexternal nigrid
198923021520   0    3    Oarcplot
198923021526   5    71   Oarcedit
198923021530   0    1    ORENAME NIGRID NIG30
198923021533   3    72   OPOLYGRID NISLAND GR10.SVF
198923021536   3    85   OGRIDPOLY GR10.SVF NI10 662795 680175 10 10
```



Lineage metadata enabled audit of data and processing
at Southern California Edison

9 visits with SCE's GIS Lab's technical staff in 1992, collected:

1. Descriptions of 14 data processing projects
2. Metadata for data sources that were acquired and imported into the enterprise GIS database for the projects
3. Processing log files for the projects



Lineage metadata enabled audit of data and processing at Southern California Edison

1. Descriptions of 14 data processing projects

   …for 7 corporate divisions were examined:

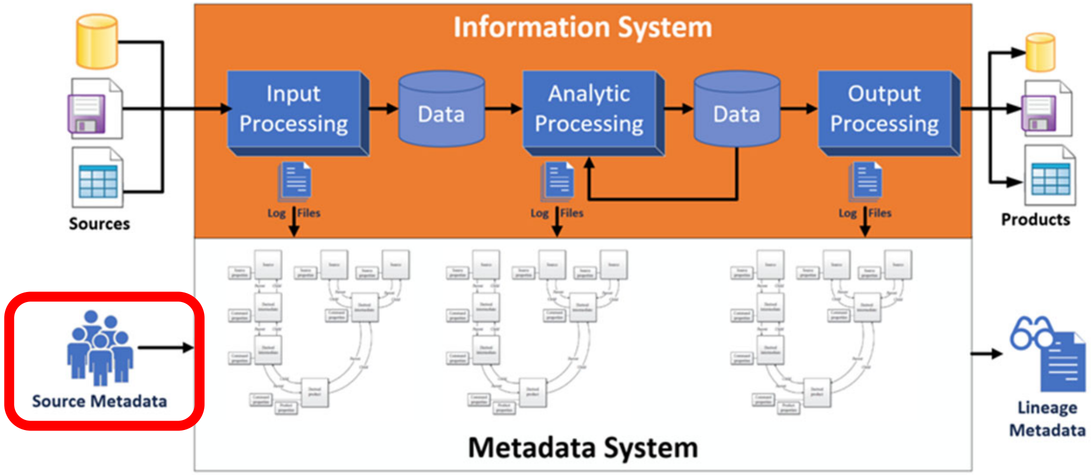
   - Customer Service
   - Engineering
   - Environmental Research
   - Information Services
   - Power Generation
   - Project Development
   - Sewer & Hydrologic Engineering

Project	Output	Deliverable
1	1 map	Spatial distribution of SCE substations relative to important features
2	5 maps	SCE's Service Territory and its various features
3	1 map	SCE's Service Territory and various features
4	1 map	Areas in Redlands CA near power lines containing sensitive species
5	1 map	Areas in Victorville CA near transmission lines containing sensitive species
6	1 map	Route of proposed pipeline from Mandalay facility to Ormond Beach facility
7	data file	Locations of historic sites in Redlands CA
8	database	Land use information for species habitat study
9	1 map	Land use, street network, elevation contours in areas around microwave stations
10	Map	Land use and street network reference map of Ormond Beach area
11	21 maps data file	3 maps each for 7 dam/reservoir sites in SCE Territory; Data file of calculated terrain units for use in hydrologic modeling project
12	database	Environmental site suitability models for locating artificial reef to mitigate impact of San Onofre Nuclear Generation Station as requirement of operation permit
13	1 map	SCE Service Territory's relationships between switching and intermediate processing centers
14	2 maps	Congressional boundaries and demographic data

Linage metadata enabled audit of data and processing
at Southern California Edison

2. Identified data acquired from internal and external sources and collected metadata on these data

- Entity types (“features”) and attribute content
- Format
- Area covered
- Scale and spatial resolution
- Spatial coordinate system
- Spatial projection
- Supplying agency
- Original source organization
- Original publication date
- Production source date
- Responsible staff member
- Statement of data quality

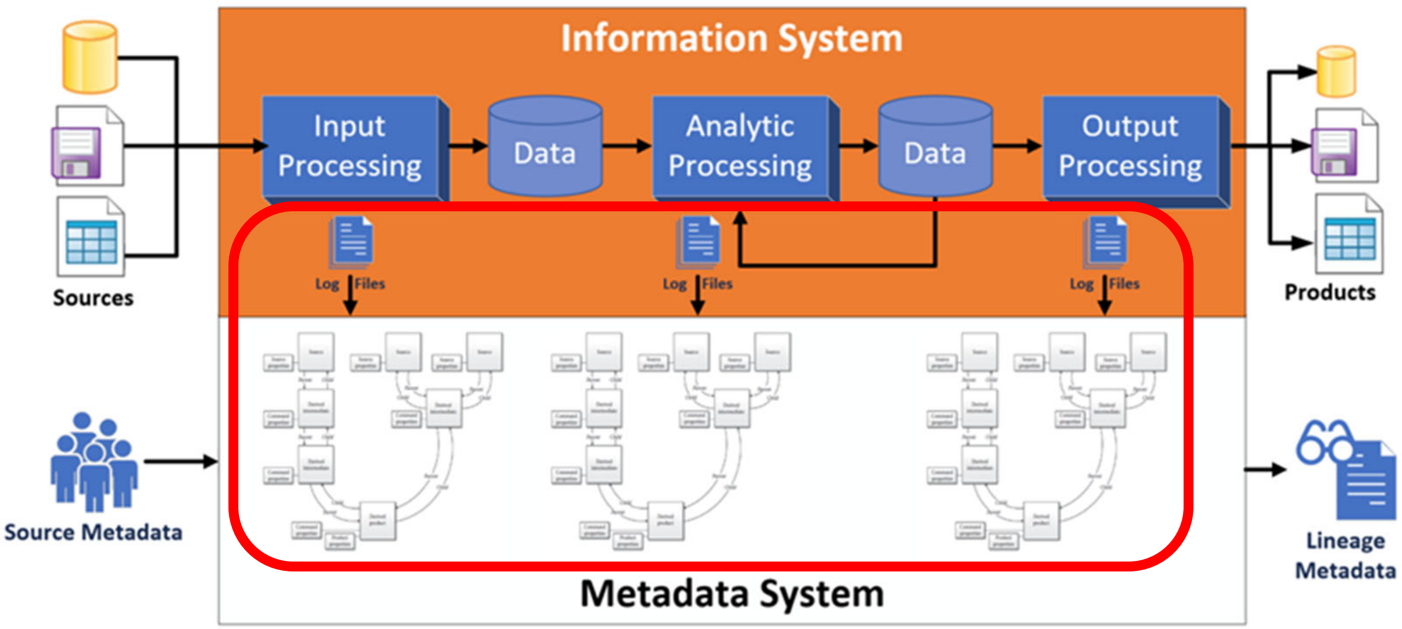


Metadata enabled audit of data and processing

at Southern California Edison

3. Processing log files obtained for each of the 14 projects

Reverse engineer lineage metadata from the log files



Metadata enabled audit of GIS data and processing

at Southern California Edison

GIS Lab analysts identified 54 data files input into the Information System to support their projects, obtained from:

- Internal client department
- Other internal departments
- California state agencies
- Outside consultants

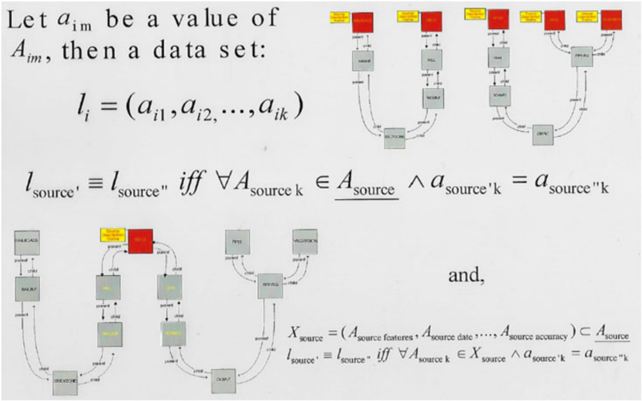
Log processing identified 806 datasets referenced in the log files :

- 487 source datasets (i.e. lacking child links pointing to inputs)
- 319 derived datasets

Metadata enabled audit of GIS data and processing
at Southern California Edison

Next step… would have focused on use of metadata analysis to identify **commonalities and differences** in:

1. Source data usage
2. Analytical processing logic



Let a_{im} be a value of A_{im}, then a data set:

$$l_i = (a_{i1}, a_{i2,} \ldots, a_{ik})$$

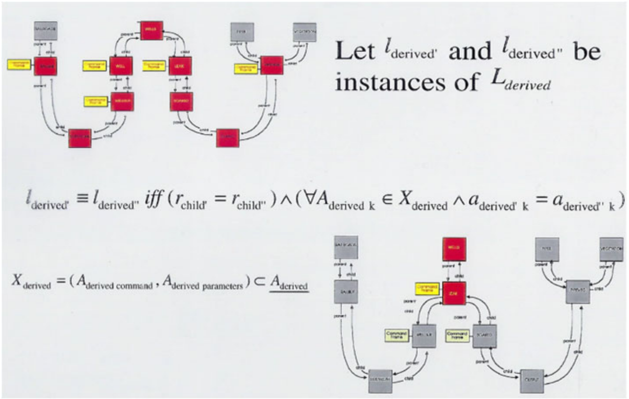
$$l_{\text{source}'} \equiv l_{\text{source}''} \ \mathit{iff} \ \forall A_{\text{source k}} \in \underline{A_{\text{source}}} \wedge a_{\text{source'k}} = a_{\text{source''k}}$$

and,

$$X_{\text{source}} = (A_{\text{source features}}, A_{\text{source date}}, \ldots, A_{\text{source accuracy}}) \subset \underline{A_{\text{source}}}$$

$$l_{\text{source}'} \equiv l_{\text{source}''} \ \mathit{iff} \ \forall A_{\text{source k}} \in X_{\text{source}} \wedge a_{\text{source'k}} = a_{\text{source''k}}$$

Source equivalence testing

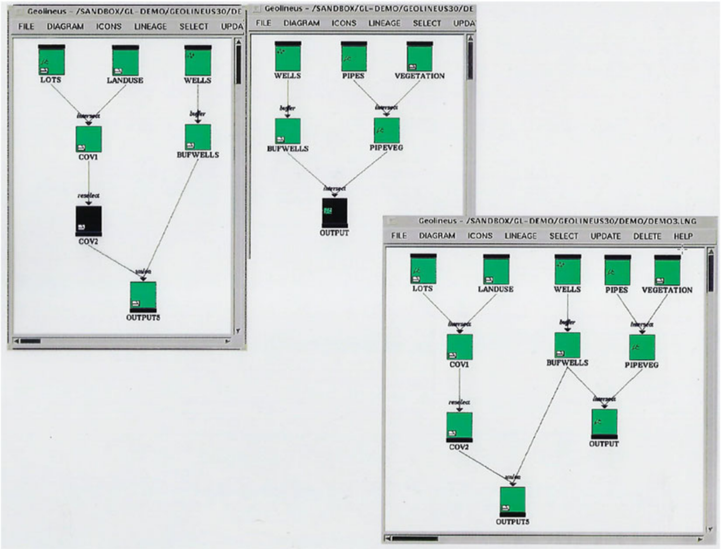


Let $l_{\text{derived}'}$ and $l_{\text{derived}''}$ be instances of $L_{derived}$

$$l_{\text{derived}'} \equiv l_{\text{derived}''} \ \mathit{iff} \ (r_{\text{child}'} = r_{\text{child}''}) \wedge (\forall A_{\text{derived k}} \in X_{\text{derived}} \wedge a_{\text{derived' k}} = a_{\text{derived'' k}})$$

$$X_{\text{derived}} = (A_{\text{derived command}}, A_{\text{derived parameters}}) \subset \underline{A_{\text{derived}}}$$

Derived equivalence testing


Geolineus - /SANDBOX/GL-DEMO/GEOLINEUS30/DE
FILE DIAGRAM ICONS LINEAGE SELECT UPDA
LOTS
LANDUSE
WELLS
intersect
buffer
COV1
BUFWELLS
reselect
COV2
union
OUTPUTS
Geolineus - /SANDBOX/GL-DEMO/GEOLINEUS30/DE
FILE DIAGRAM ICONS LINEAGE SELECT UPDA
WELLS
PIPES
VEGETATION
buffer
intersect
BUFWELLS
PIPEVEG
intersect
OUTPUT
Geolineus - /SANDBOX/GL-DEMO/GEOLINEUS30/DEMO/DEMO3.LNG
FILE DIAGRAM ICONS LINEAGE SELECT UPDATE DELETE HELP
LOTS
LANDUSE
WELLS
PIPES
VEGETATION
intersect
buffer
intersect
COV1
BUFWELLS
PIPEVEG
reselect
intersect
COV2
OUTPUT
union
OUTPUTS

Metadata enabled audit of GIS data and processing
at Southern California Edison

But… findings:

1. Much metadata for documenting the data sources were missing…

   - GIS Lab Technical Staff analysts were unable to remember much about the data they had used in earlier projects

   - Of the 54 data files used as input to the GIS database:
     - 89% were of unknown Spatial Projections
     - 79% were of unknown Original Publication Dates
     - 70% were of unknown Scales and Spatial Resolutions
     - 68% were from unknown Original Source Organizations
     - 43% contained attributes and spatial data assumed “fit for use” but untested

Metadata enabled audit of GIS data and processing at Southern California Edison

Findings:

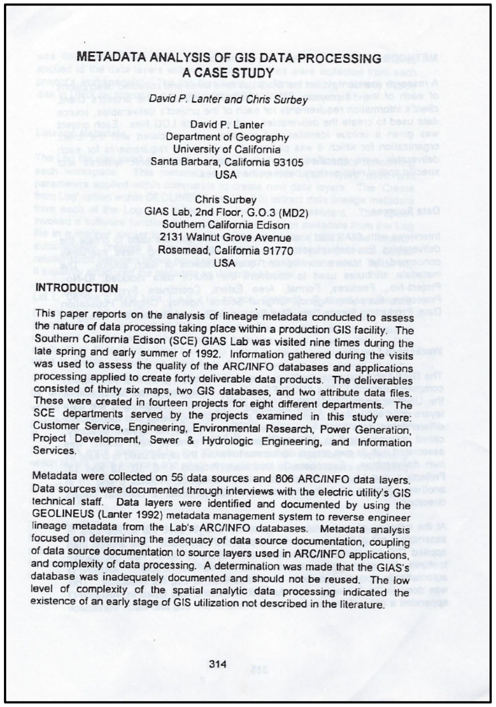
2. Lack of naming conventions for identifying primary data source files and source datasets once they were imported into the Information System
   - For example,
     - "TER" used as mnemonic device to name datasets after import:
       - 5 datasets in Project 1: TERBND, TER.MRK, TERMRK1, TERMRK2, and TERMERK3
       - 3 datasets in Project 2: TERRITORY, SCE-TERR, SCE-TERR2
       - Information Analysts could not differentiate them

Utility company only had one service territory boundary, there were 8 different versions of it. Without taking the itme to visually inspect and compare the actual data – it was not clear what, if any, significant differences existed among the versions

Metadata enabled audit of GIS data and processing at Southern California Edison

Recommendation:

- GIS Lab's *"…database was inadequately documented and should not be reused."*



METADATA ANALYSIS OF GIS DATA PROCESSING
A CASE STUDY

David P. Lanter and Chris Surbey

David P. Lanter
Department of Geography
University of California
Santa Barbara, California 93105
USA

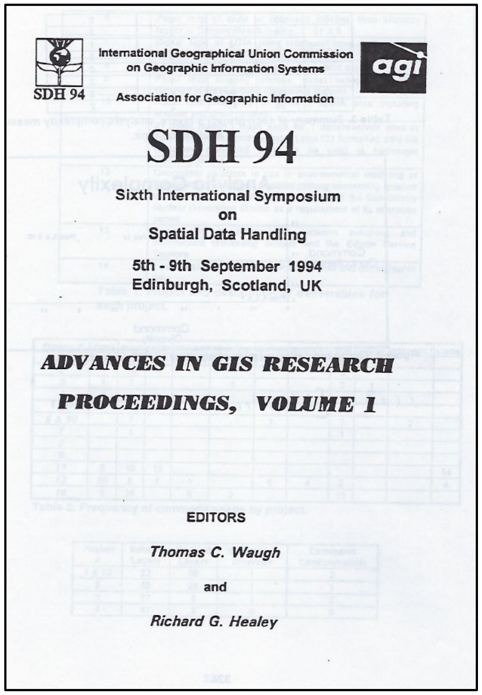
Chris Surbey
GIAS Lab, 2nd Floor, G.O.3 (MD2)
Southern California Edison
2131 Walnut Grove Avenue
Rosemead, California 91770
USA

INTRODUCTION

This paper reports on the analysis of lineage metadata conducted to assess the nature of data processing taking place within a production GIS facility. The Southern California Edison (SCE) GIAS Lab was visited nine times during the late spring and early summer of 1992. Information gathered during the visits was used to assess the quality of the ARC/INFO databases and applications processing applied to create forty deliverable data products. The deliverables consisted of thirty six maps, two GIS databases, and two attribute data files. These were created in fourteen projects for eight different departments. The SCE departments served by the projects examined in this study were: Customer Service, Engineering, Environmental Research, Power Generation, Project Development, Sewer & Hydrologic Engineering, and Information Services.

Metadata were collected on 56 data sources and 806 ARC/INFO data layers. Data sources were documented through interviews with the electric utility's GIS technical staff. Data layers were identified and documented by using the GEOLINEUS (Lanter 1992) metadata management system to reverse engineer lineage metadata from the Lab's ARC/INFO databases. Metadata analysis focused on determining the adequacy of data source documentation, coupling of data source documentation to source layers used in ARC/INFO applications, and complexity of data processing. A determination was made that the GIAS's database was inadequately documented and should not be reused. The low level of complexity of the spatial analytic data processing indicated the existence of an early stage of GIS utilization not described in the literature.

314



SDH 94

International Geographical Union Commission
on Geographic Information Systems

agi

Association for Geographic Information

SDH 94

Sixth International Symposium
on
Spatial Data Handling

5th - 9th September 1994
Edinburgh, Scotland, UK

ADVANCES IN GIS RESEARCH

PROCEEDINGS, VOLUME 1

EDITORS

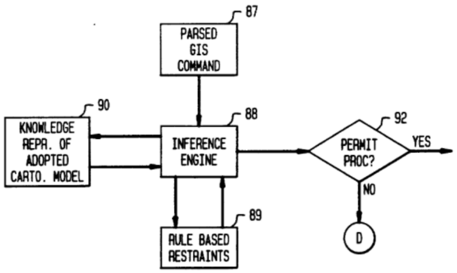
Thomas C. Waugh

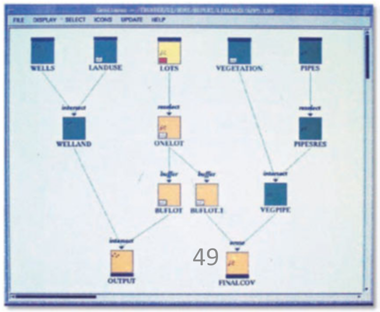
and

Richard G. Healey

Conclusion: Data lineage metadata can help information systems meet key data privacy by design requirements, including:

- Enabling data subjects to access, review and rectify their personal data
- Enable data subjects to withdraw given consent with effect for the future by:
  a. Blocking access to their personal data?
  b. Constraining processing and usage of their personal data?
  c. Erasing their personal data?
- Blocking and restricting personal data obtained for one purpose from being processed for other purposes not compatible with the original purpose



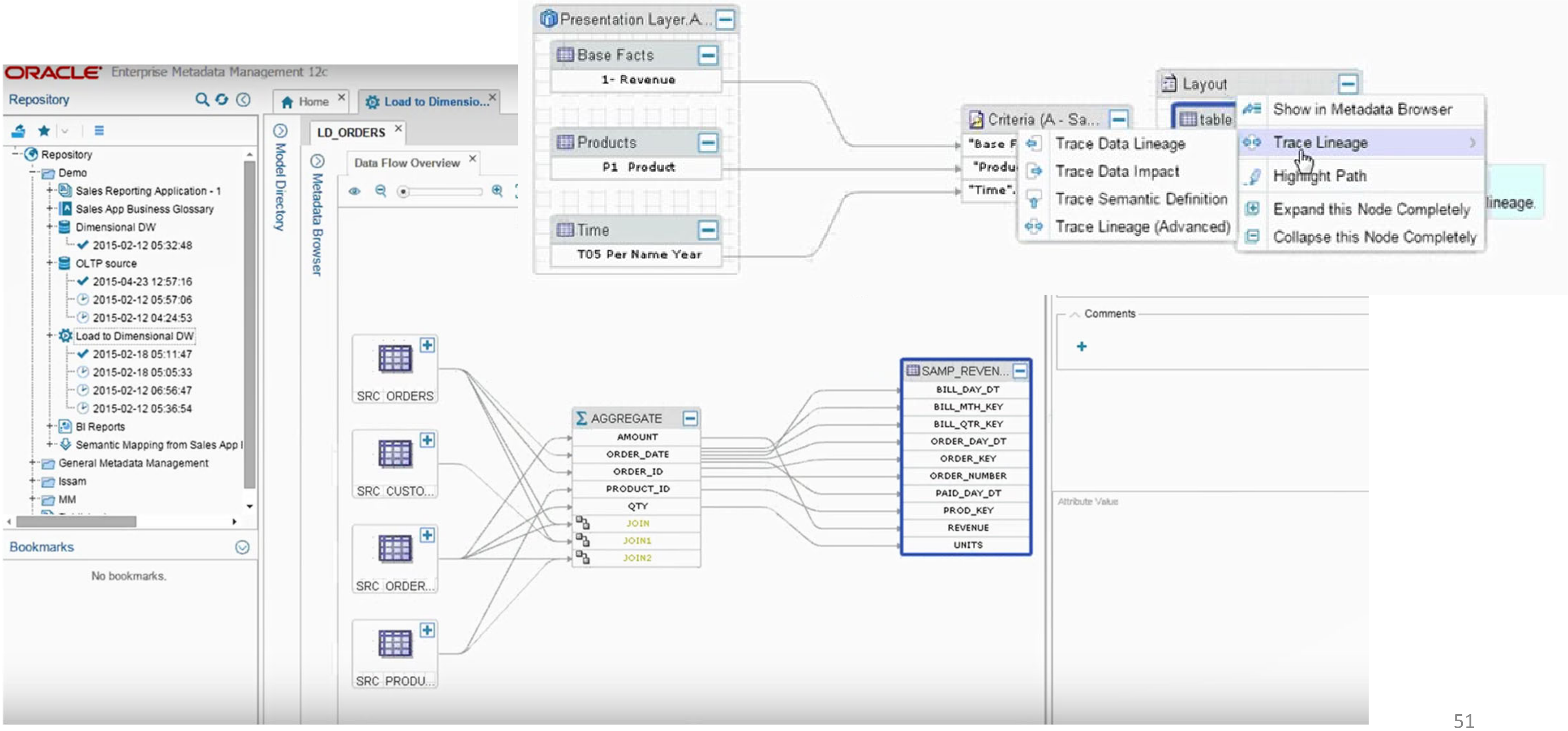


Conclusion:

Data lineage metadata can be used to help information system developers meet key data protection by design requirements:

1. Data subjects have **right to access, review and rectify** their personal data
2. Data subjects have the **right to withdraw given consent** with effect for the future and
   - Block access
   - Constrain processing and use
   - Erase their personal data
3. Personal **data obtained for one purpose must not be processed for other purposes** not compatible with the original purpose

Outlook: Commercial database management systems are beginning to include lineage metadata capabilities for tracking attribute values processed and transformed among relational database tables ...



Agenda

- ✓ Data protection by design
- System Security Plan
  - Security control inheritance
  - Team project SSP review and discussion

Agenda

- ✓Data protection by design
- ✓Cloud computing specifications
- Security control origination
- Team project SSP progress review and discussion

Security Control Origination

Security control “inheritance” exist when

an information system or application receives protection from security controls developed, implemented, assessed, authorized, and monitored by entities other than those responsible for the system or application

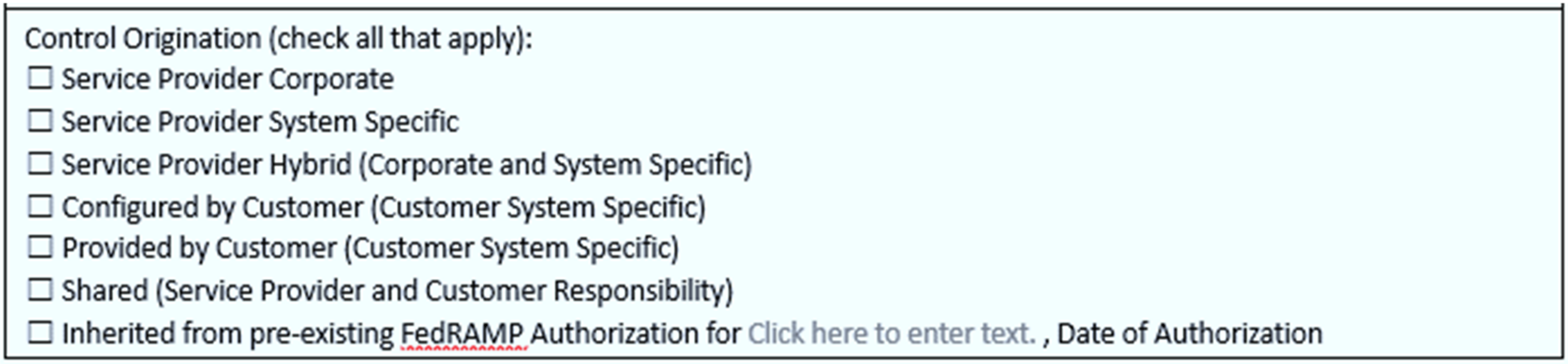
NIST SP 800-53 Revision 4

Control Origination

Many of the controls needed to protect organizational information systems are inheritable by other systems, e.g.

- Security awareness training
- Incident response plans
- Physical access to facilities
- Rules of behavior
- Public Key Infrastructure [PKI]
- Authorized secure standard configurations for clients/servers
- Access control systems
- Boundary protection
- Cross-domain solutions

Control Origination



Control Origination (check all that apply):

□ Service Provider Corporate

□ Service Provider System Specific

□ Service Provider Hybrid (Corporate and System Specific)

□ Configured by Customer (Customer System Specific)

□ Provided by Customer (Customer System Specific)

□ Shared (Service Provider and Customer Responsibility)

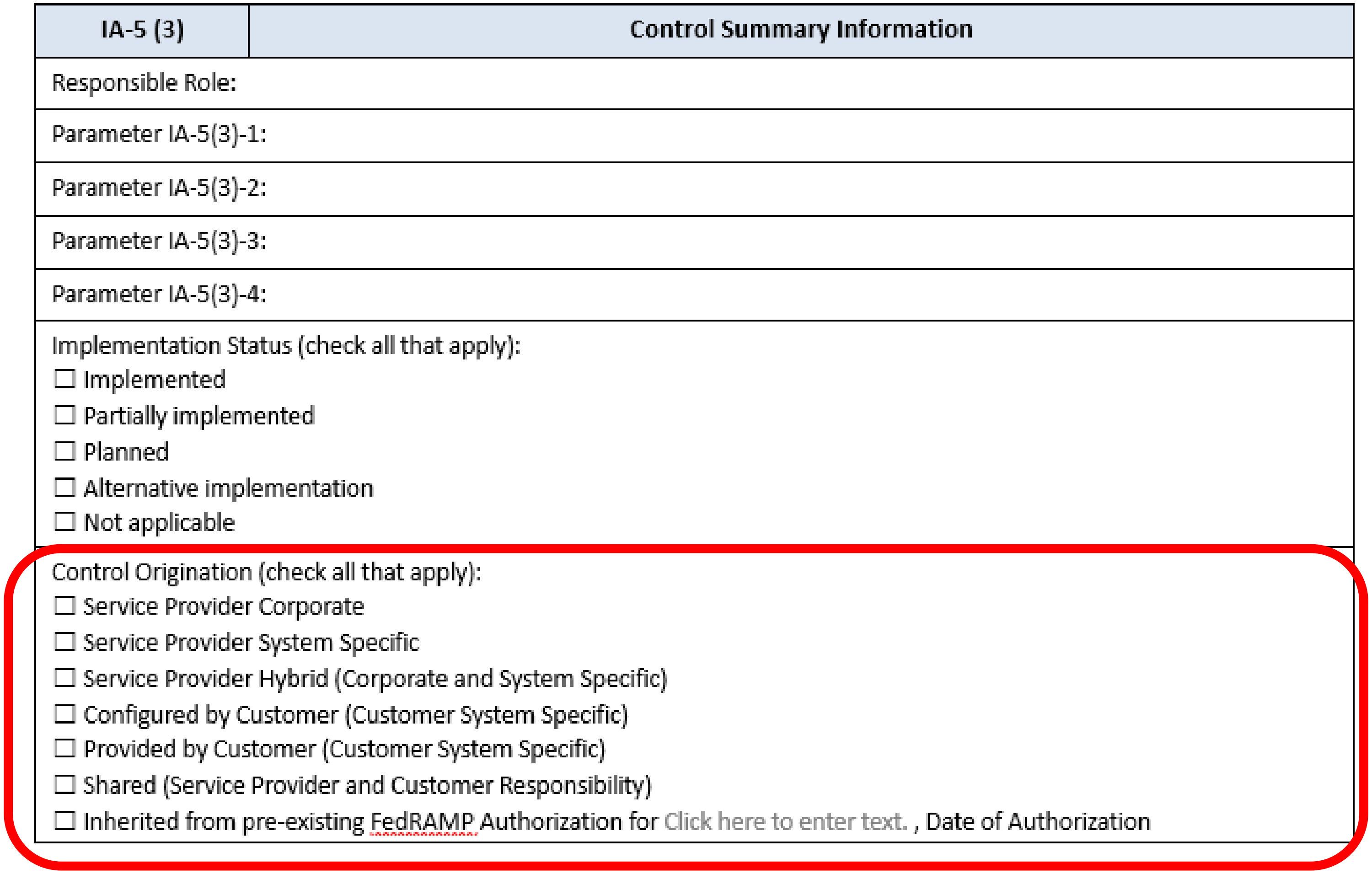
□ Inherited from pre-existing FedRAMP Authorization for Click here to enter text. , Date of Authorization

- Indicate what sections of the security control are inherited and provide a description of what is inherited
- If a entire control is inherited, it must be clear to the Assessor what is inherited
- The writer does not need to describe how the leveraged service is performing the particular function
  - That detail is found in the SSP of the leveraged system from which the control is inherited

If a policy has been published and is referenced as is the basis for the implementation of the inherited security control, make sure that published document is provided as an attachment, or a supporting artifact with the SSP when submitted for FedRAMP review

https://www.fedramp.gov/weekly-tips-cues-february-15-2017/

Control Origination



IA-5 (3)	Control Summary Information
Responsible Role:	
Parameter IA-5(3)-1:	
Parameter IA-5(3)-2:	
Parameter IA-5(3)-3:	
Parameter IA-5(3)-4:	
Implementation Status (check all that apply): ☐ Implemented ☐ Partially implemented ☐ Planned ☐ Alternative implementation ☐ Not applicable	
Control Origination (check all that apply): ☐ Service Provider Corporate ☐ Service Provider System Specific ☐ Service Provider Hybrid (Corporate and System Specific) ☐ Configured by Customer (Customer System Specific) ☐ Provided by Customer (Customer System Specific) ☐ Shared (Service Provider and Customer Responsibility) ☐ Inherited from pre-existing FedRAMP Authorization for Click here to enter text. , Date of Authorization	

Agenda

- ✓Data protection by design
- ✓System Security Plan
  - ✓Cloud computing specifications
  - ✓Security control inheritance
  - Team project SSP review and discussion