

Unit #11

Data Protection

MIS5214

Agenda

- Data protection by design
- System Security Plan
 - Cloud computing specifications
 - 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**”

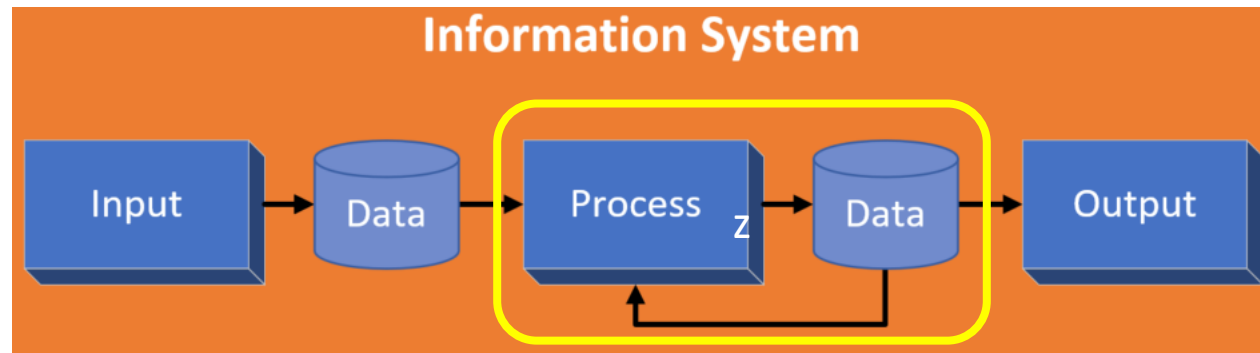
Danezis, G. et al. (2014) “Privacy and Data Protection by Design”,
European Union Agency for Network and Information Security (ENISA)

Some challenging data protection requirements may be solved with techniques presented in this webinar...

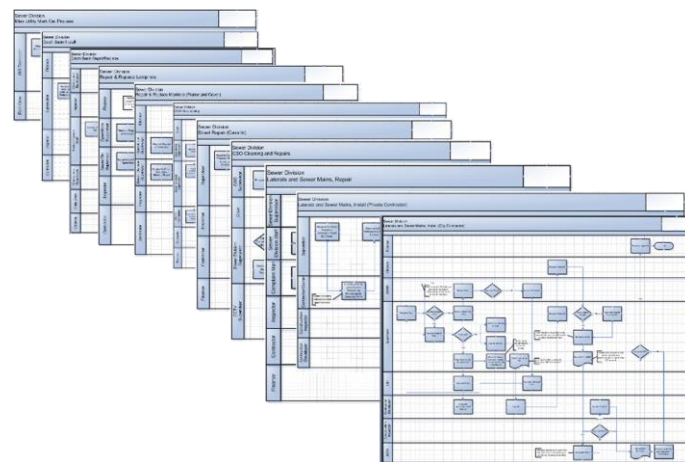
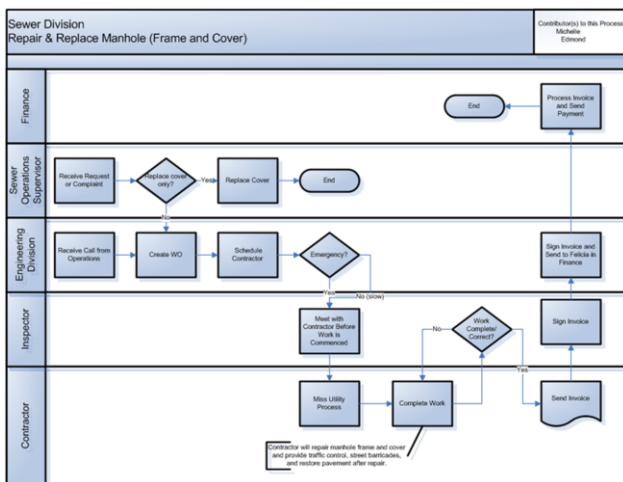
1. **Collection** of personal data is **fully avoided or minimized** at the earliest stage of processing
2. Data subjects give **specific, informed and explicit consent** to the processing of their data
3. Data subjects have **right to access, review and rectify** their personal data
4. Data subjects 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...

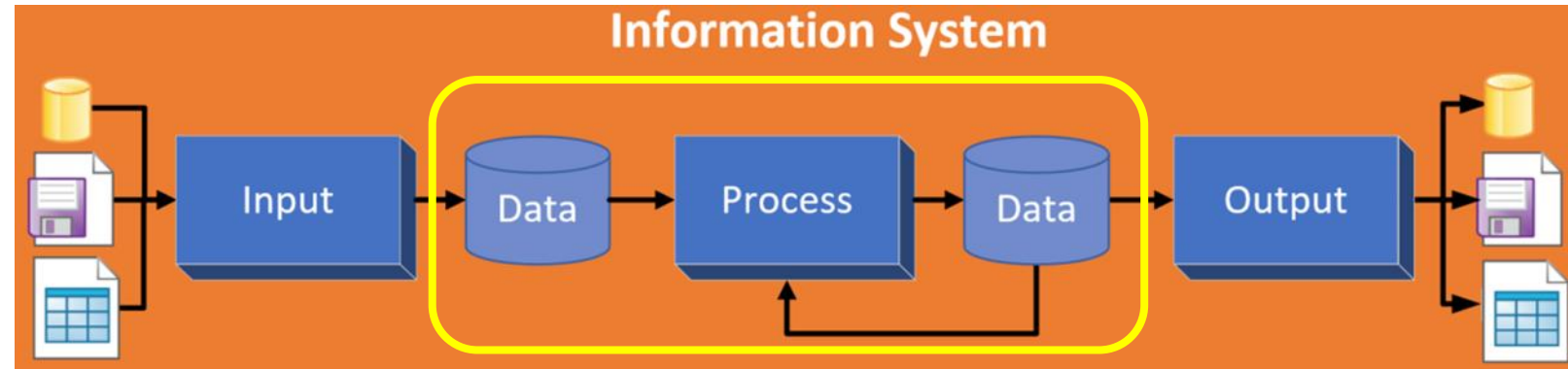


Regardless of application, there is reliance on data processing workflows to produce and use information

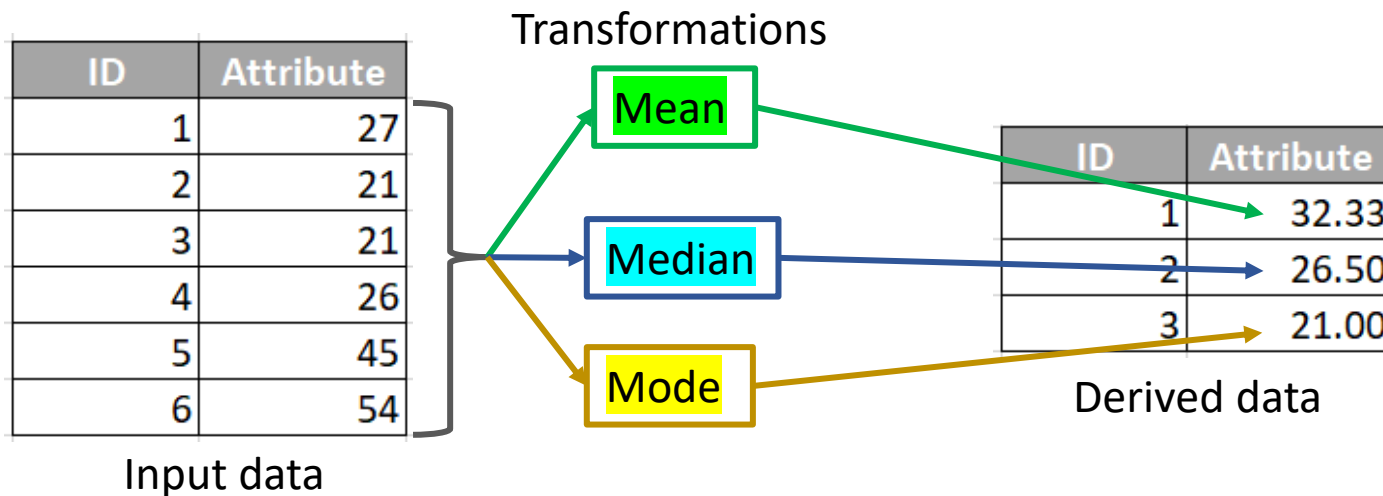


Data processing often transforms existing data into new data, which is a double-edged sword...

➤ *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*



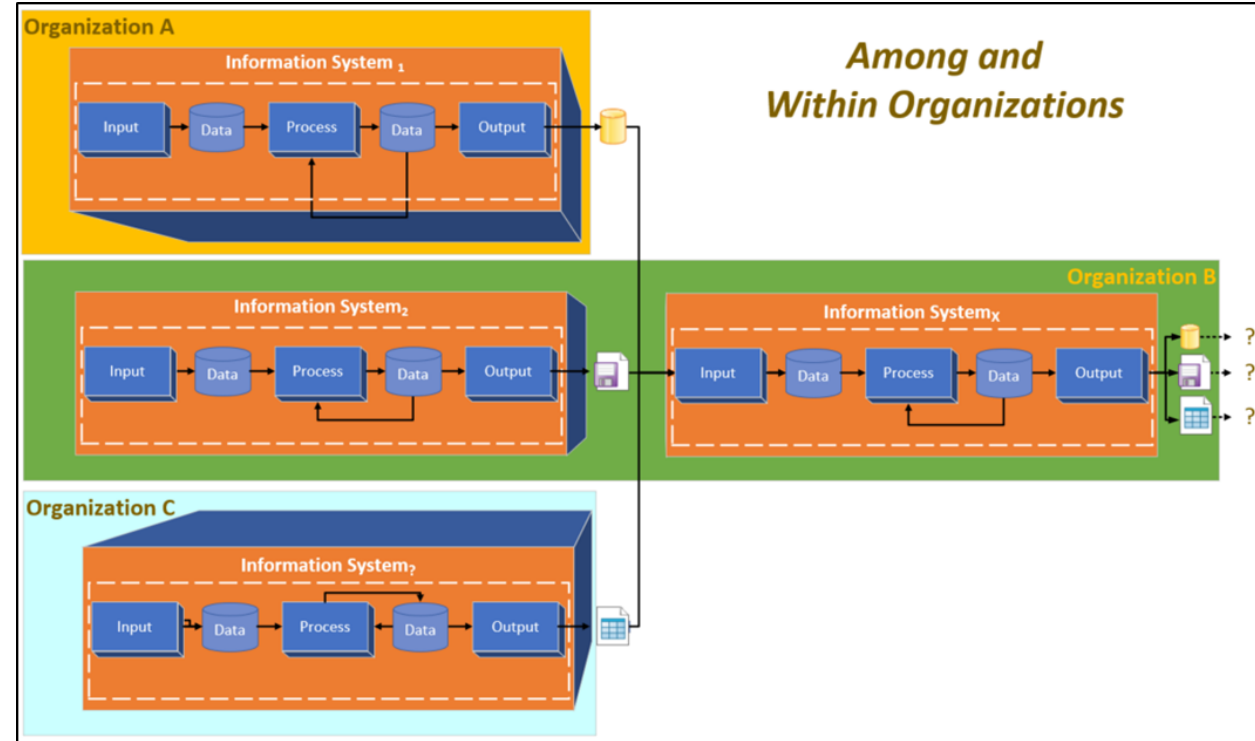
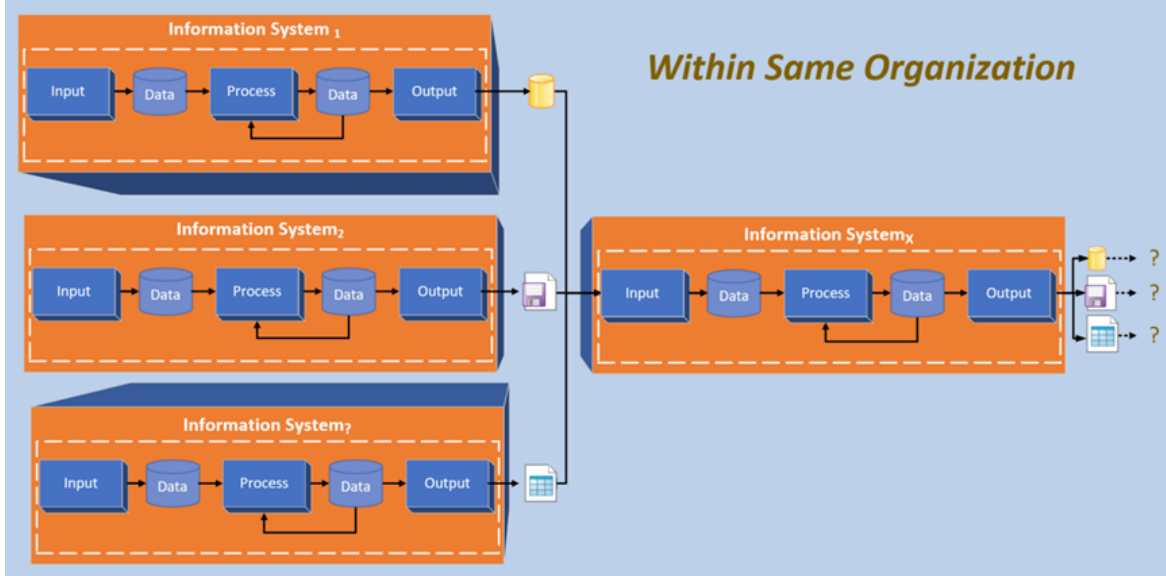
Evaluating & judging data's "fitness for use"

- **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

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

Durand-Ruel, Paris, August 23, 1872 [1];
Catholina Lambert, New Jersey;
Lambert sale, American Art Association, Plaza Hotel, New York, NY,
February 21, 1916 until February 24, 1916, no. 67;
Durand-Ruel, Paris, until at least 1930;
purchased by Simon Bauer, Paris, by June 1936 [2];
anonymous sale, Parke-Bernet Galleries, Inc., February 25, 1970, no. 19 [3];
Sam Salz, Inc., New York, NY;
purchased by Museum, May 1971.

Notes:

[1] bought from the artist.

[2] Listed and illustrated in "List of Property Removed from France during the War 1939-1945" (no. 7114, as belonging to Simon Bauer).

[3] "Highly Important Impressionist, Post-Impressionist & Modern Paintings and Drawings", illustrated.

Standardizing Museum Provenance – David Newbury (@workergnome)

Newbury, D. (2017) "Standardizing Museum Provenance for the Twenty-First Century", from talk given at the Yale Center for British Art

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):

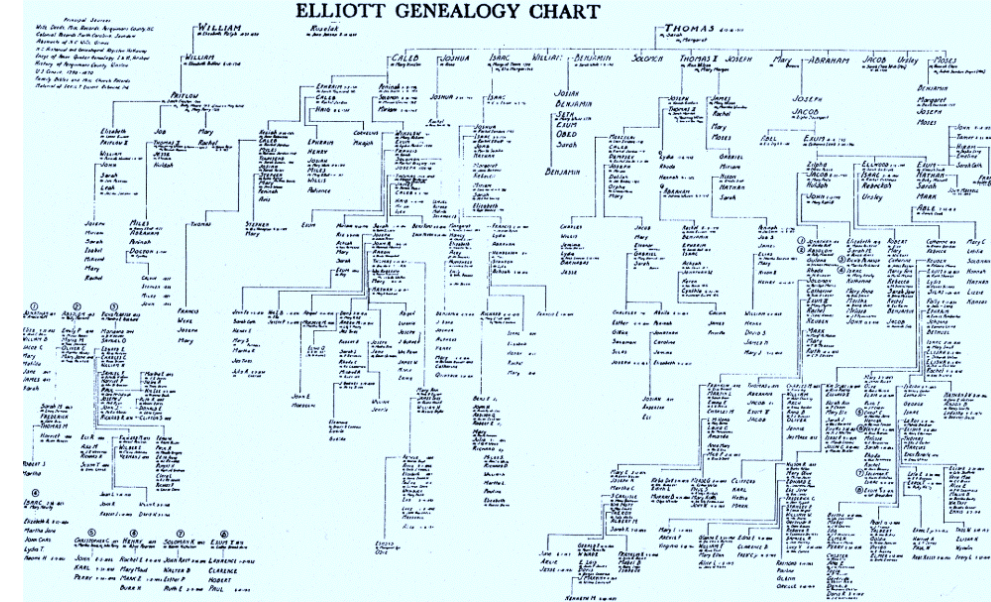
- Provenance is a record that describes entities and processes involved in producing and delivering or influencing a resource
- Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility
- Provenance assertions are contextual metadata that can become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

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

Provenance and data lineage

“Data provenance” and “data lineage” is used here interchangeably, overlooking subtle differences in their meanings


- Data provenance suggests process history
- Data lineage implies a kind of genealogy or data pedigree record relative to both
 1. Sources of data
 2. Processing applied to the sources to produce an information product



This presentation explores how data lineage metadata can aid understanding and establish trust of data

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

FGDC-STD-001-1998



NSDI
National Spatial Data Infrastructure

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 **EN ISO 19115-1**
NORME EUROPÉENNE
EUROPÄISCHE NORM

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

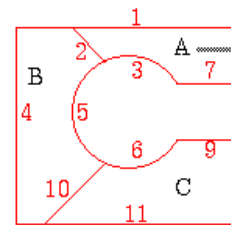
CEN-CENELEC Management Centre: Avenue Marnix 17, B-1000 Brussels

© 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

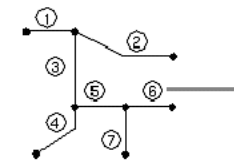
- Also organize their data in
 - Data base management systems
 - File systems



Polygon Attribute Table

Polygon	Area	Parcel Number	Land Use
A	12,001	11-115-001	R 1
B	15,775	11-115-002	R 1
C	19,136	11-115-003	R 3

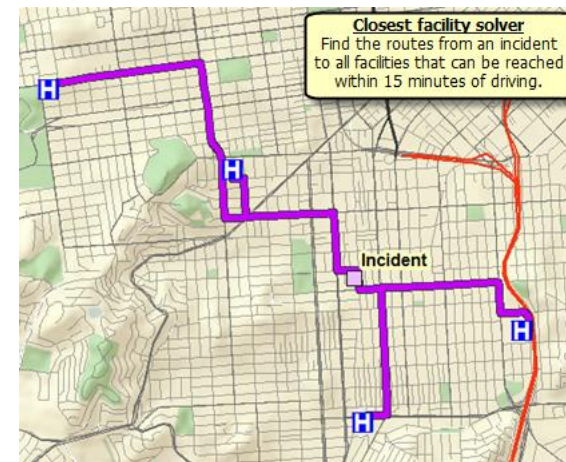
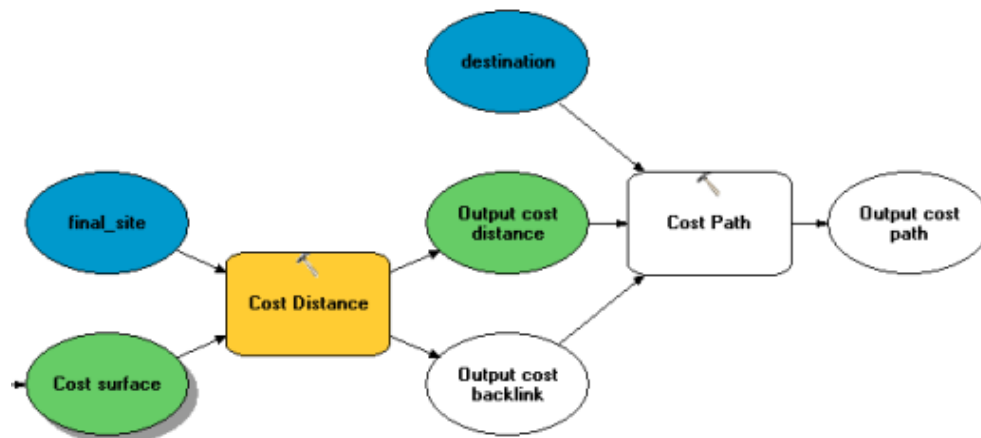
Coverage: Roads



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





National Spatial Data Infrastructure

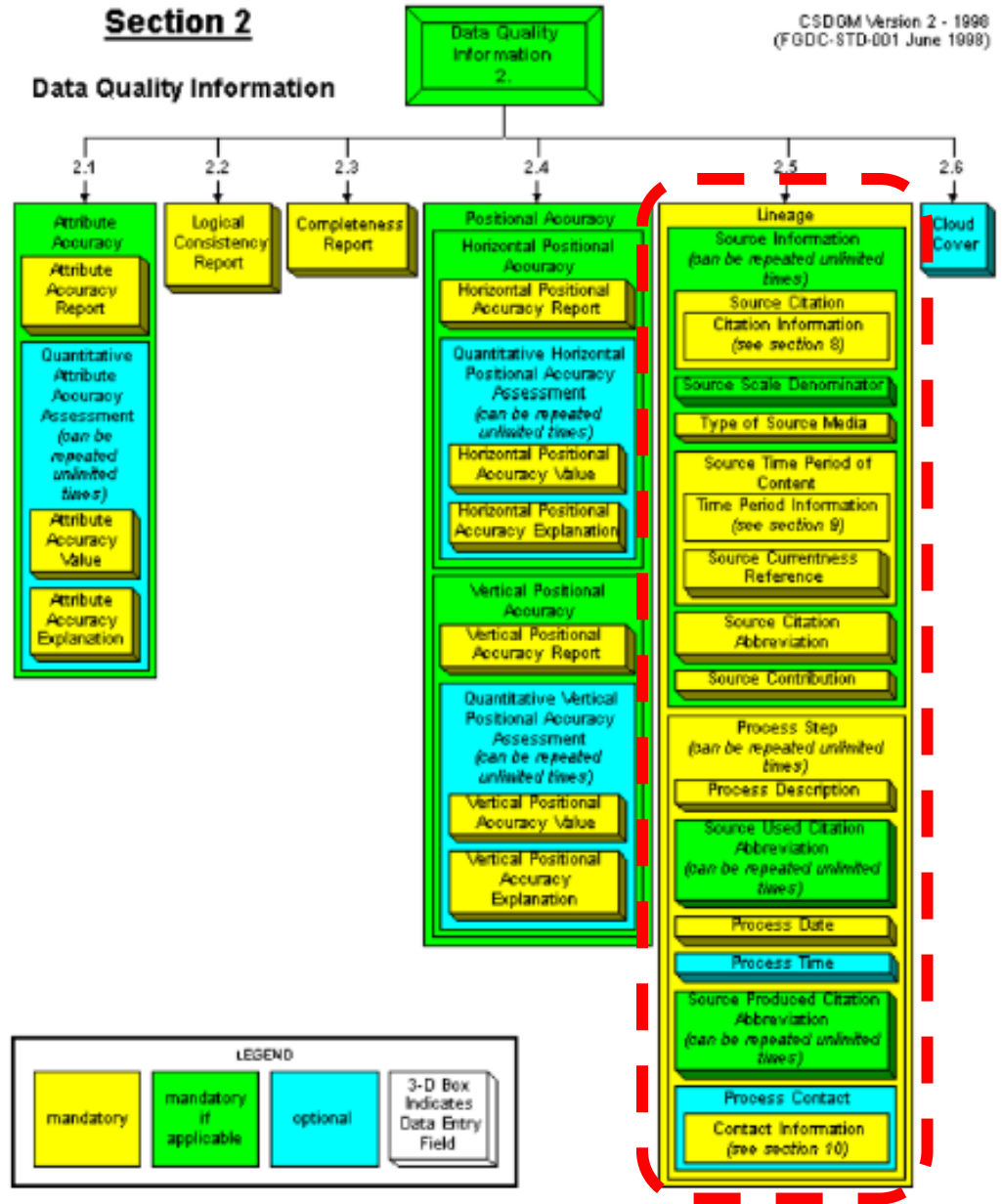
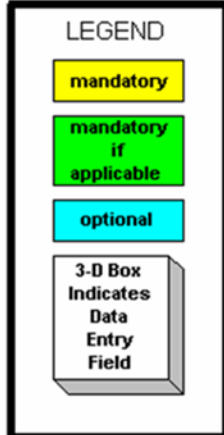
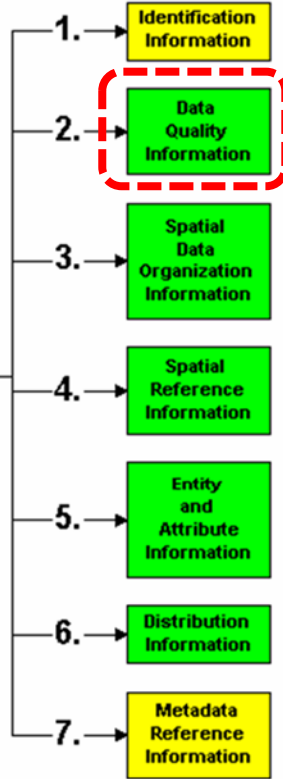
FGDC-STD-001-1998

Content Standard for Digital Geospatial Metadata

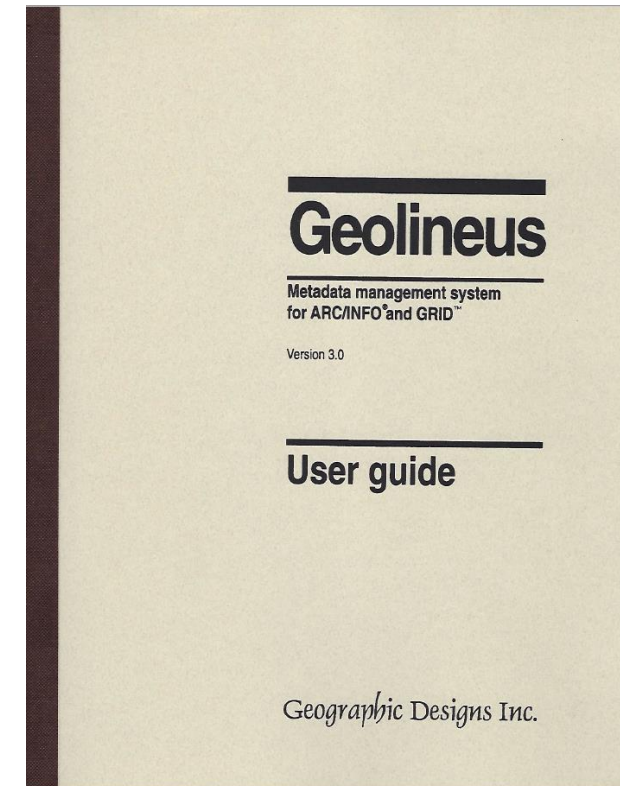
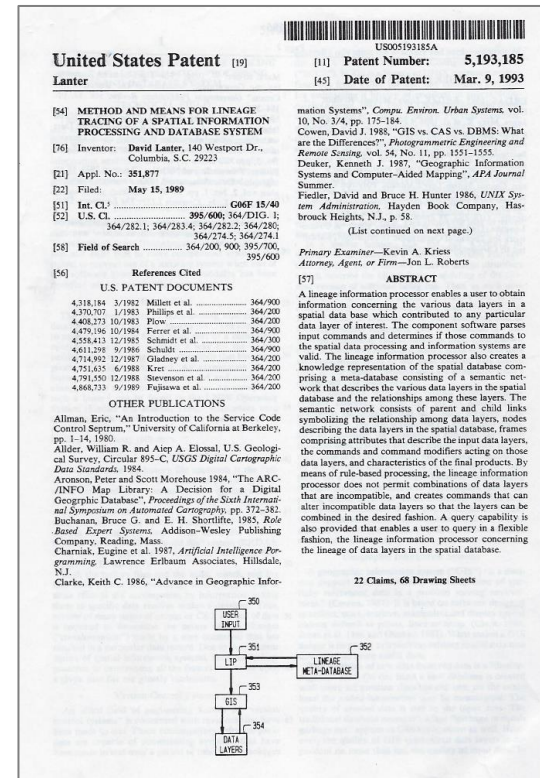
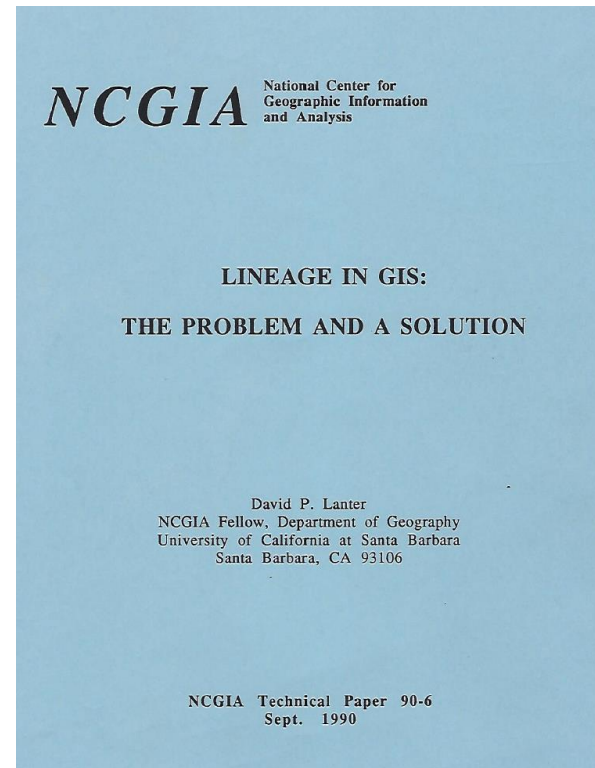
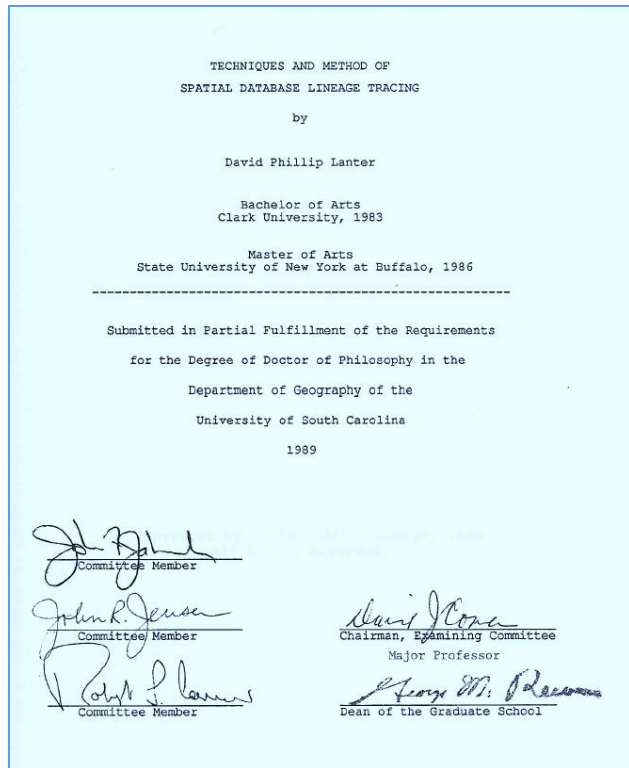
Metadata Ad Hoc Working Group
Federal Geographic Data Committee

Federal Geographic Data Committee

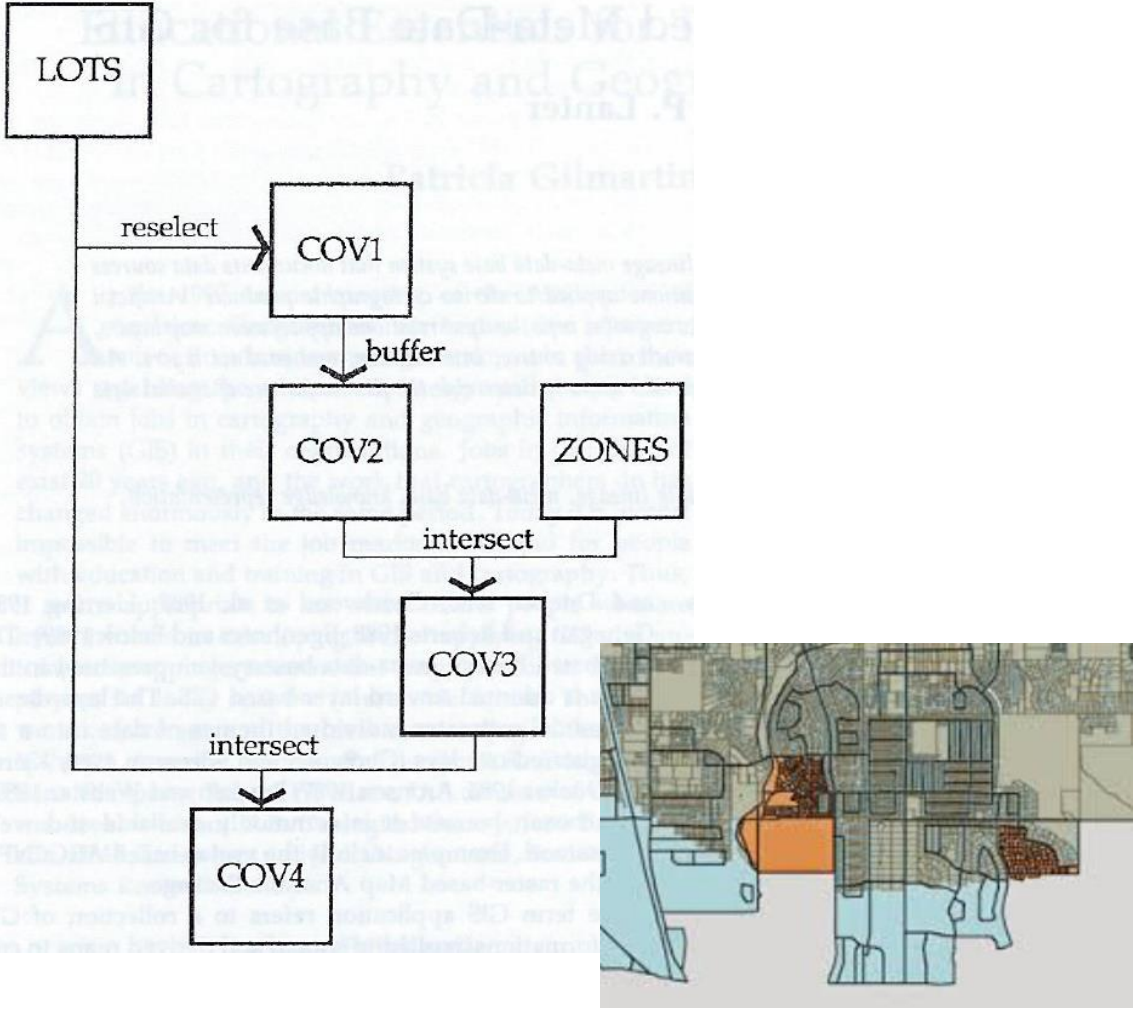
Department of Agriculture • Department of Commerce • Department of Defense • Department of Energy
Department of Housing and Urban Development • Department of the Interior • Department of State
Department of Transportation • Environmental Protection Agency
Federal Emergency Management Agency • Library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



1st Metadata Processing System focused on data lineage (provenance)



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

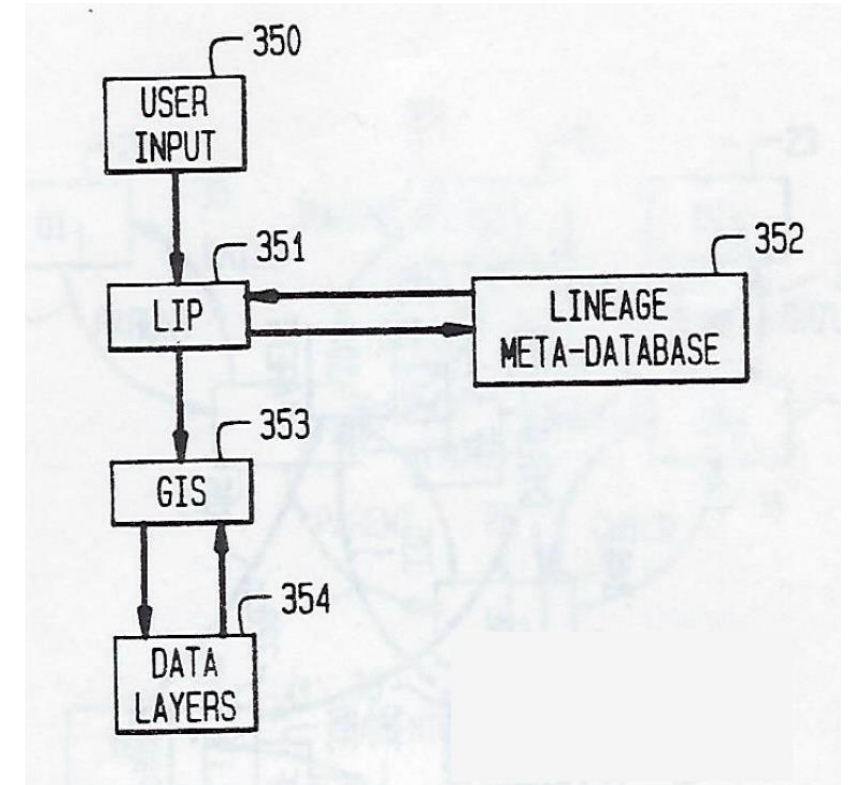
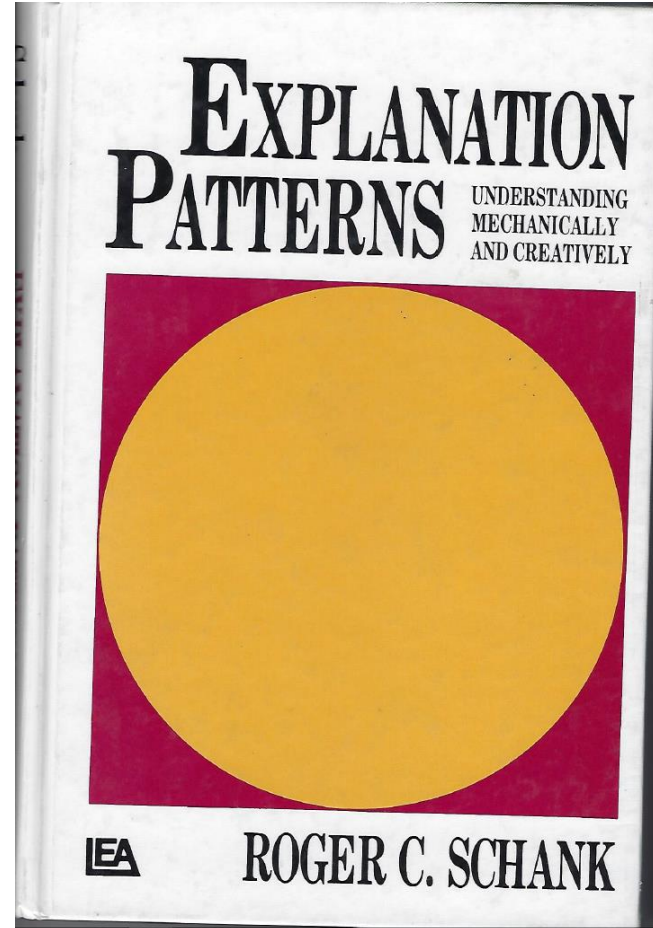
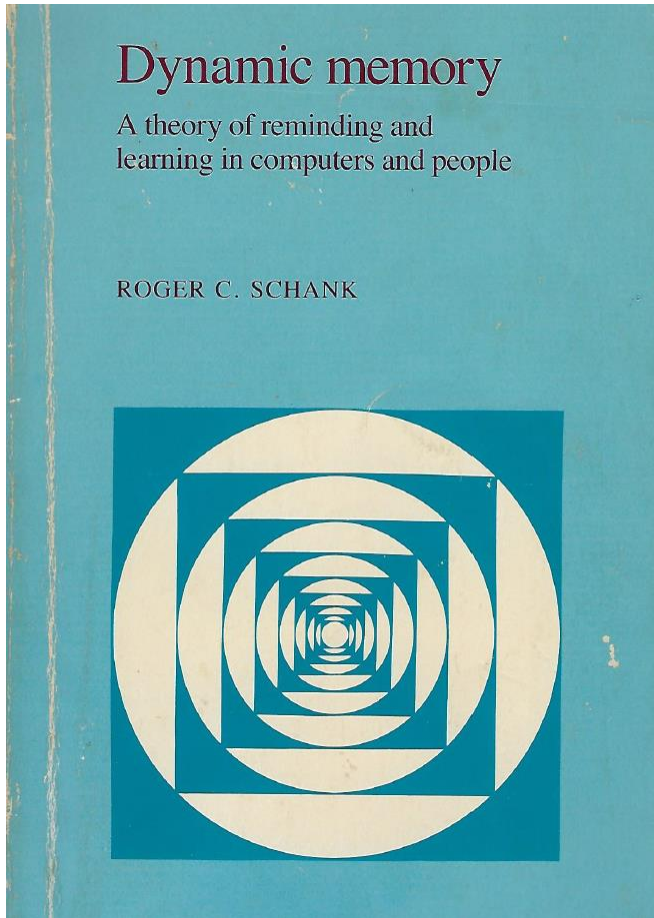


Datasets presented by the operating system after data processing concluded...

Datasets organized as files in folders

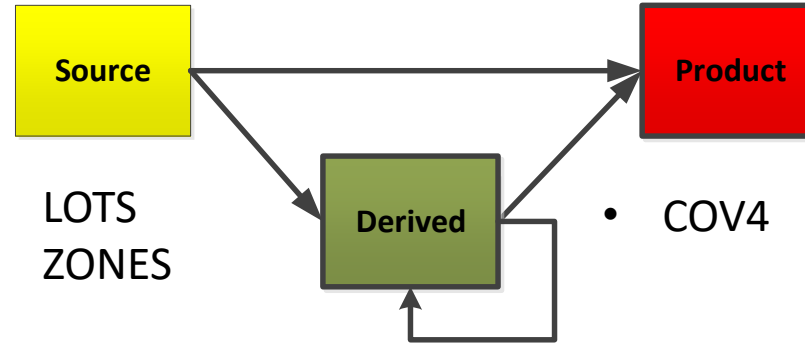
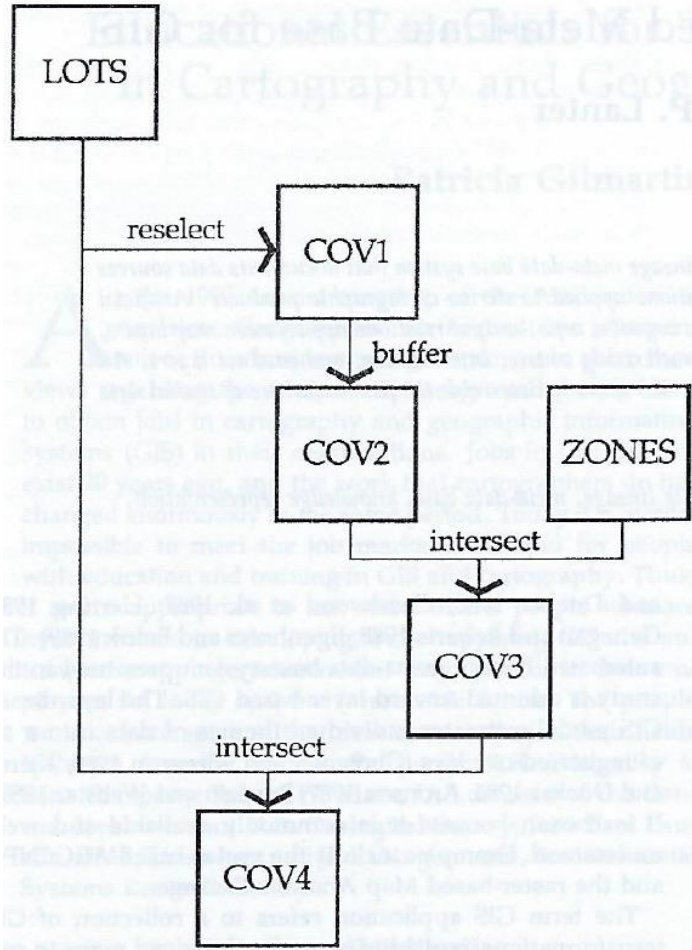
.	<DIR>	5-05-89	10:26a
..	<DIR>	5-05-89	10:26a
COV1	<DIR>	5-24-89	11:35p
LOTS	<DIR>	5-05-89	10:26a
INFO	<DIR>	5-05-89	10:26a
ZONES	<DIR>	5-05-89	10:27a
OUTPUT	<DIR>	5-05-89	10:27a
ONELOT	<DIR>	5-06-89	11:52a
DAV1	<DIR>	5-31-89	1:35p
FINAL	<DIR>	5-06-89	12:27p
COV3	<DIR>	5-24-89	11:46p
COV4	<DIR>	5-24-89	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?



LIP = Lineage Information Processor

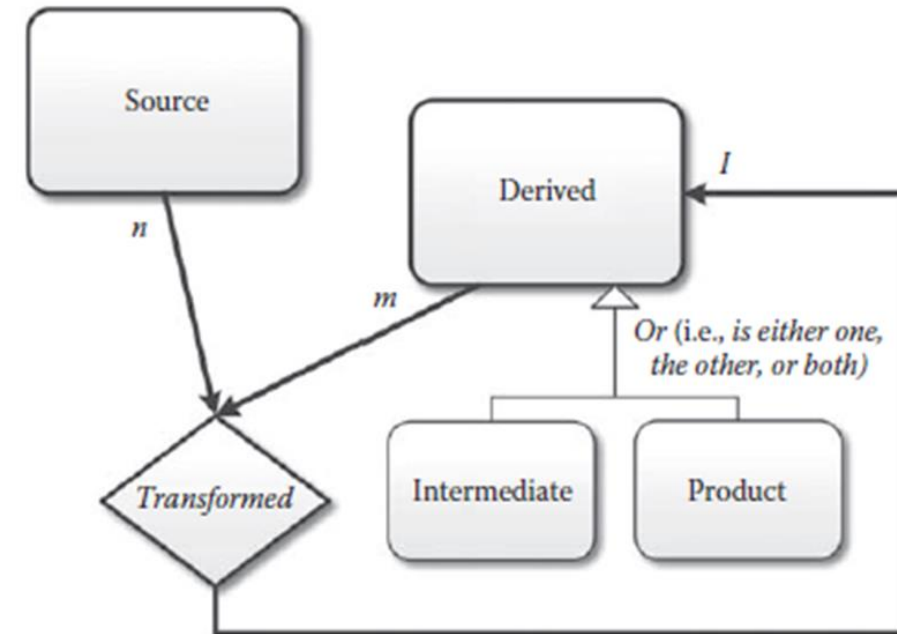
How do we understand differences among datasets created during processing applications?



- LOTS
- ZONES

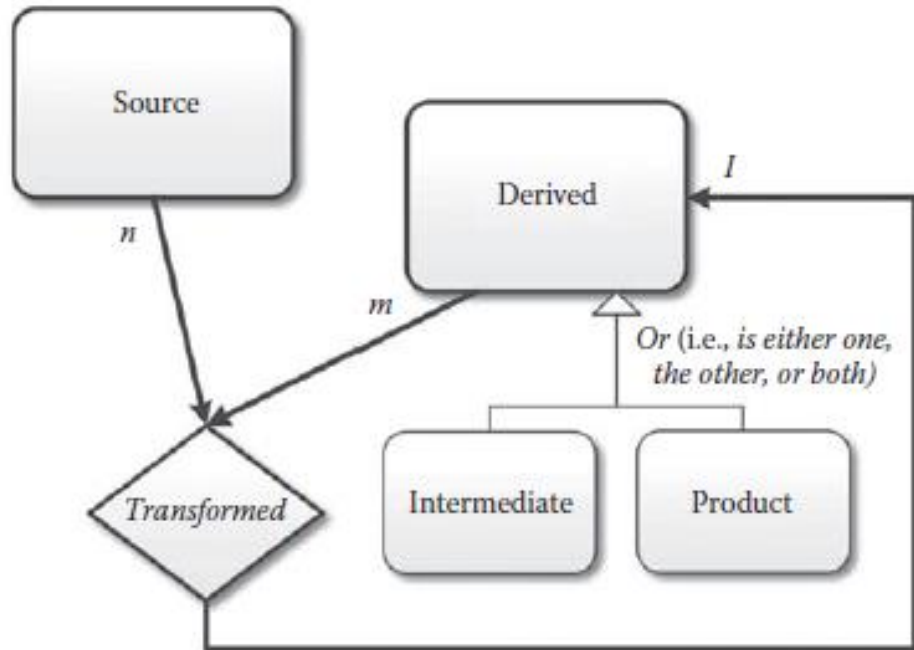
- COV4

- 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

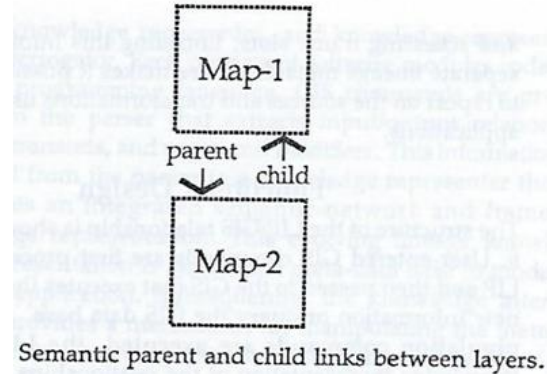
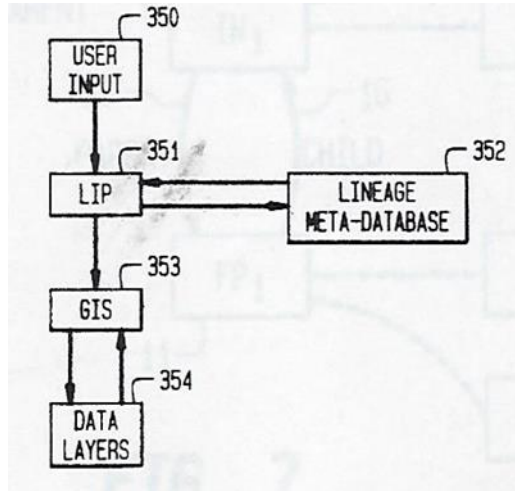


Source datasets *may contain personal data*

Derived datasets inherit this personal data from their input

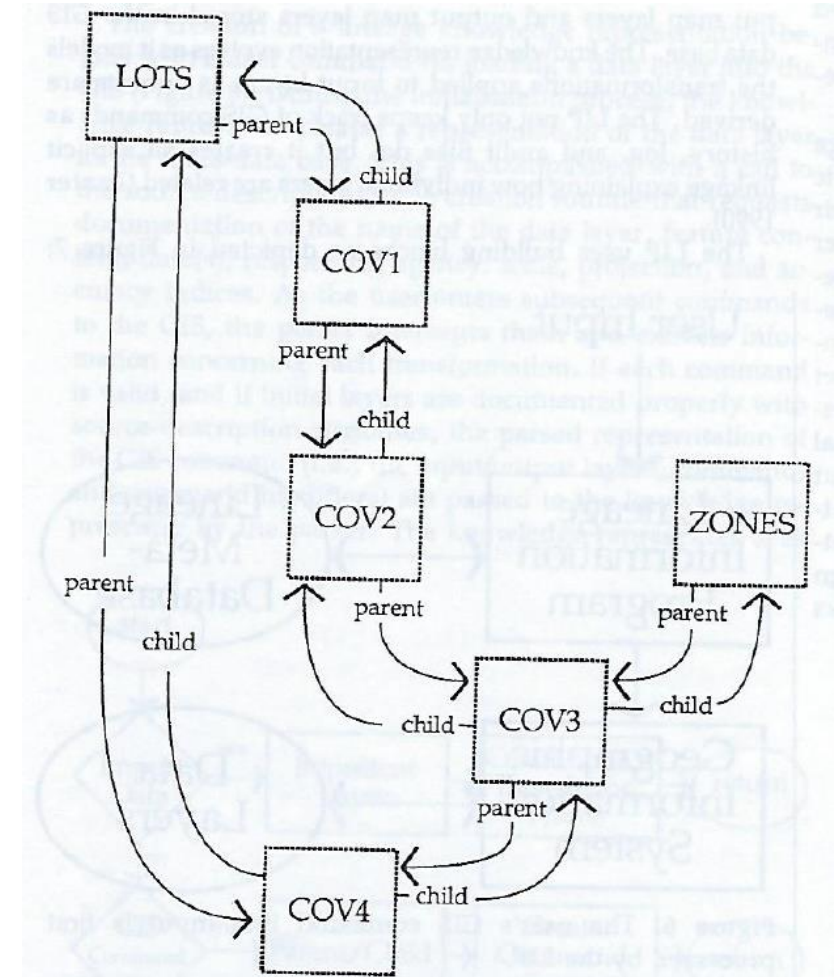
- *Using transformations such as:*
 - *Relational database joins and relates*
 - *Queries, arithmetic, statistical, spatial processing...*

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



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?***



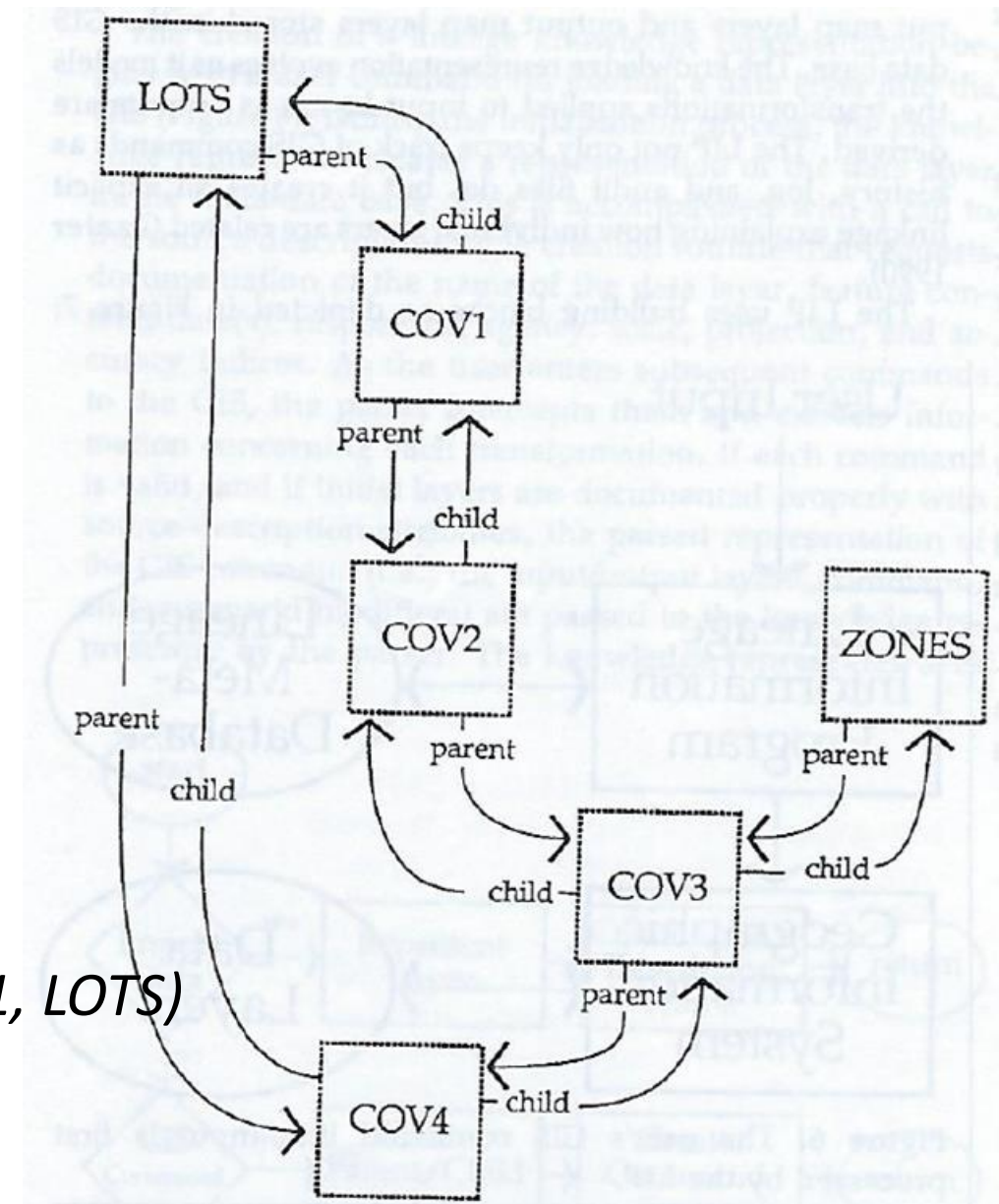
Descendants function traces parent links to identify all datasets derived from a source or other derived input dataset used within the application.

```
(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)



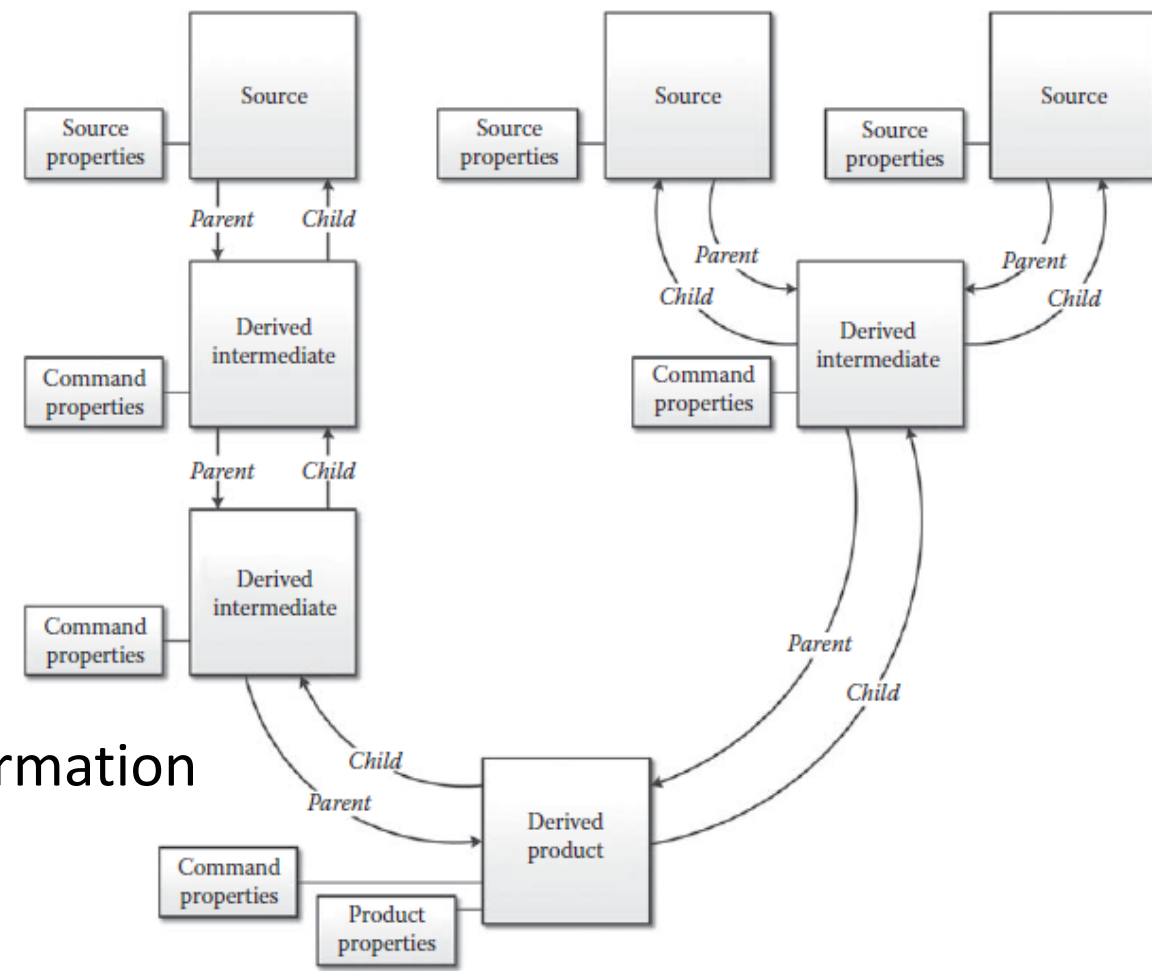
Source properties can include:

- 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
```

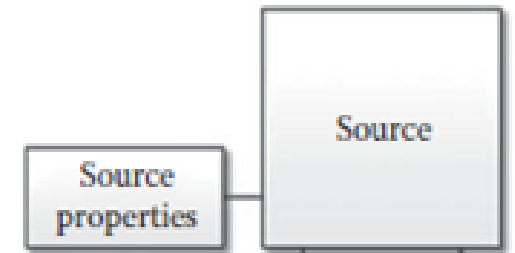
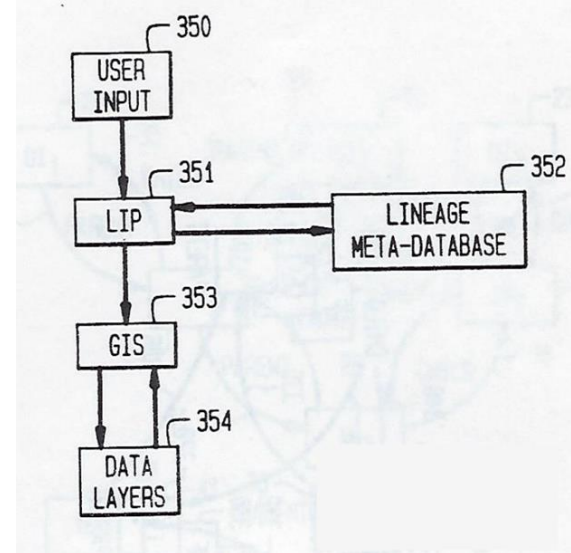
```
What is the source agency? USGS
```

```
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)

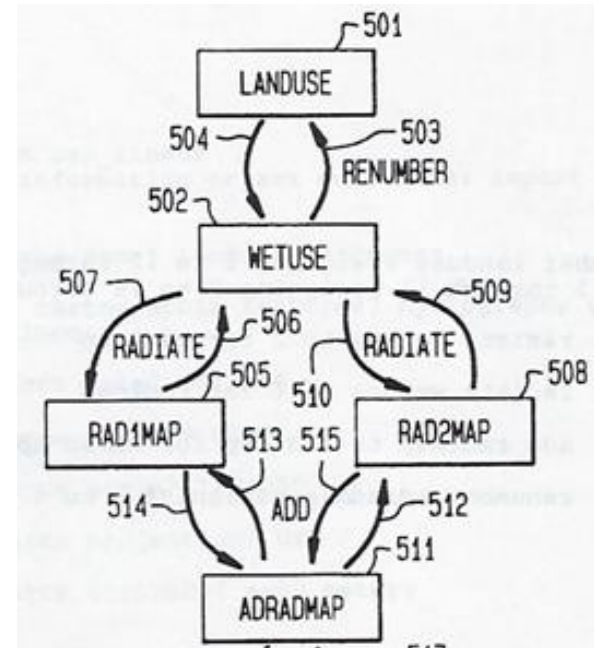
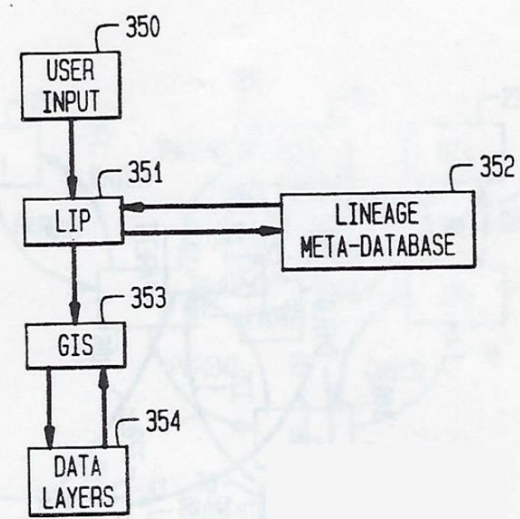
(I AM GEO_LINEUS)

(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)

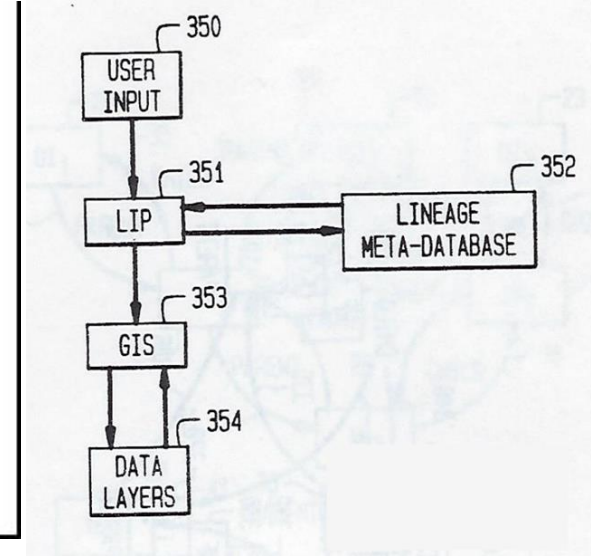
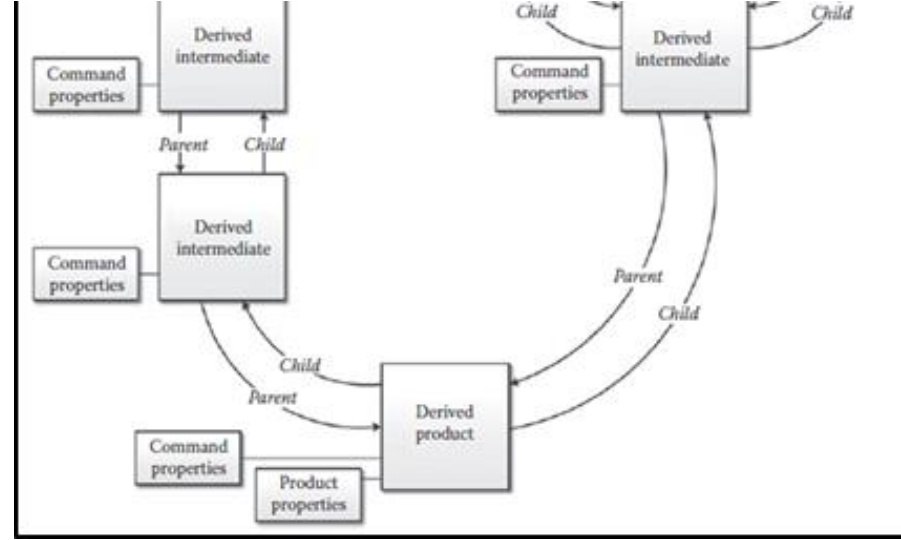
(I UNDERSTAND) (radiate wetuse to 2 for rad1map)

(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

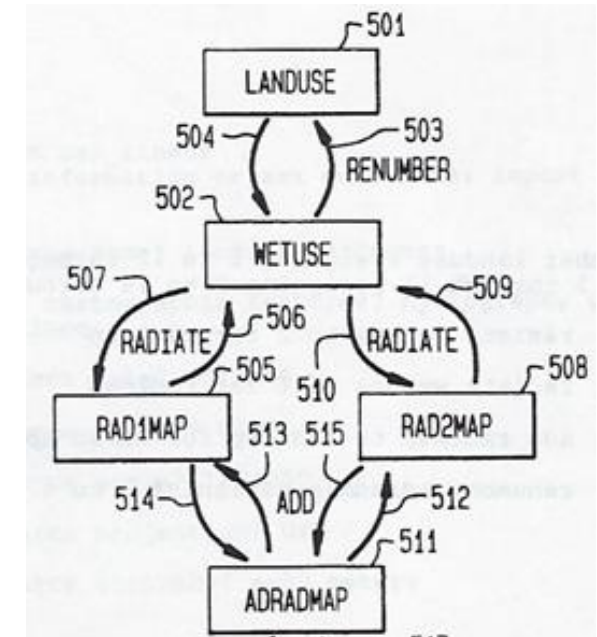
What is the product's use? Environmental protection of wetlands

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

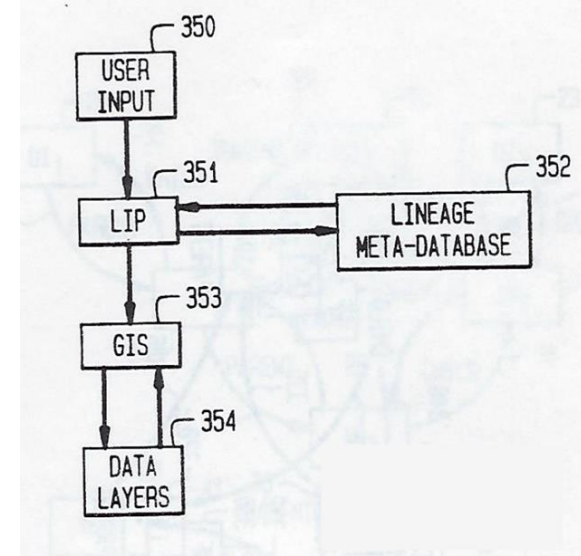
Who is responsible for the product? Diego Essinger

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

Thank You!

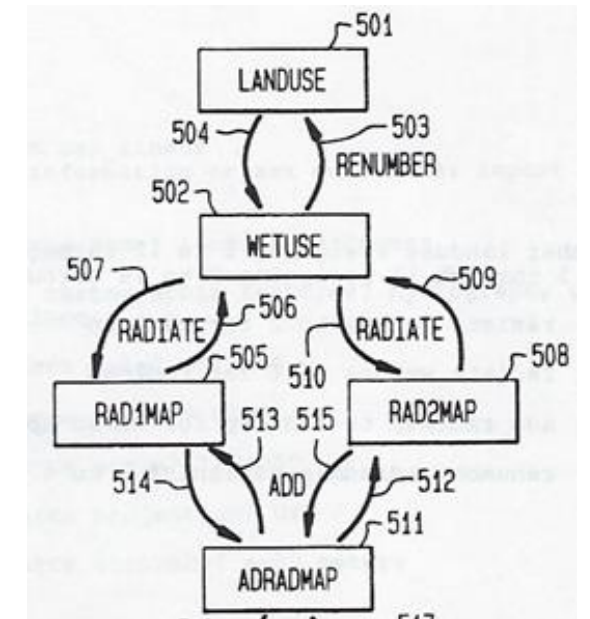


Querying metadata...

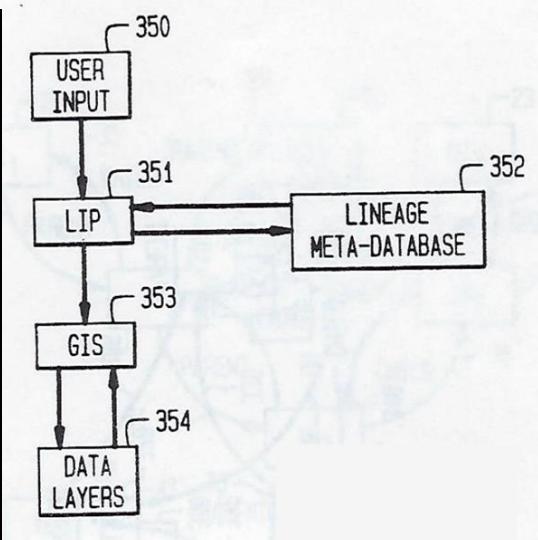
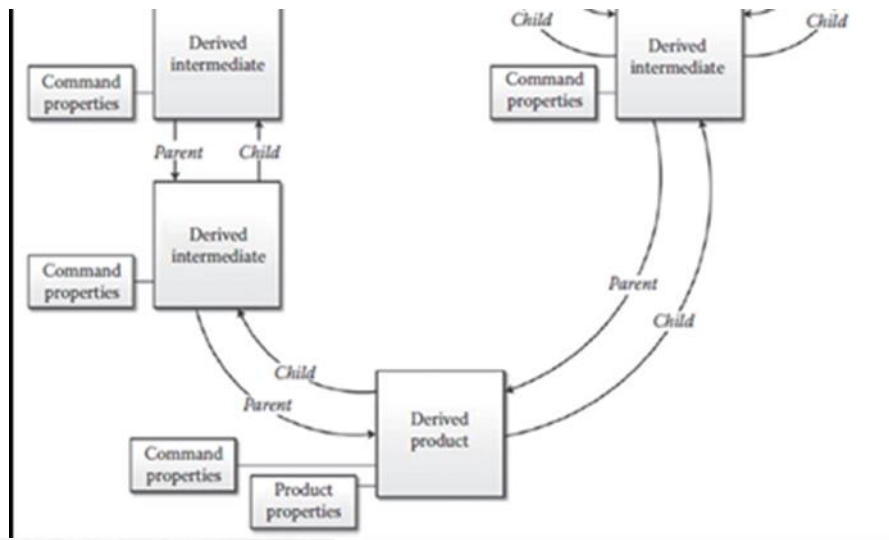


Is landuse a parent of adradmap

(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)

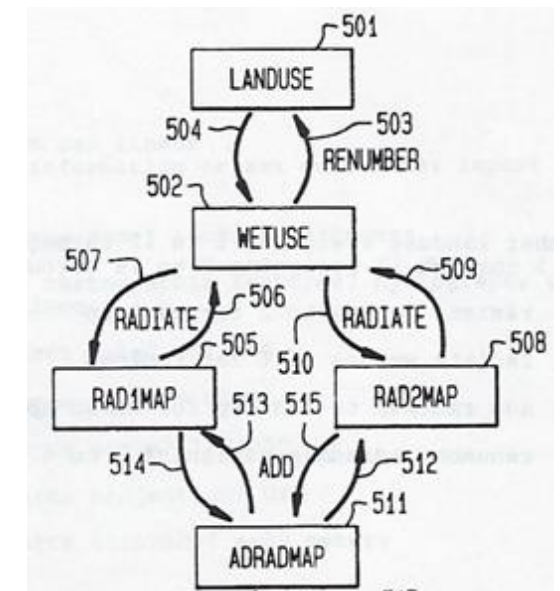
(LANDUSE IS AN ORIGINAL MAP LAYER)

(INPUT TO RAD1MAP IS WETUSE COMMAND IS RADIATE)

(INPUT TO WETUSE IS LANDUSE COMMAND IS RENUMBER)

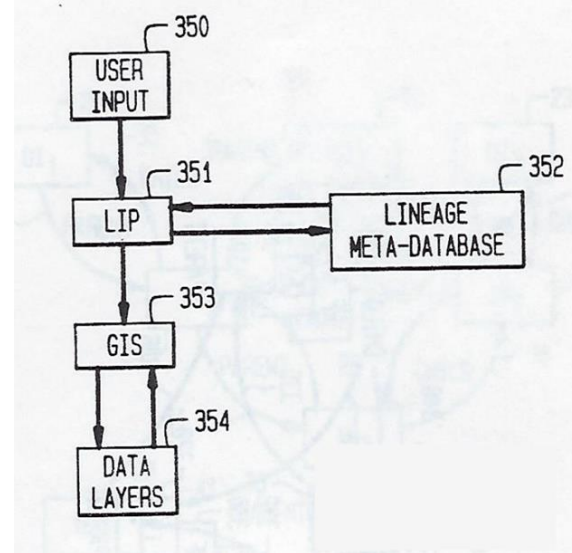
(LANDUSE IS AN ORIGINAL MAP LAYER)

+

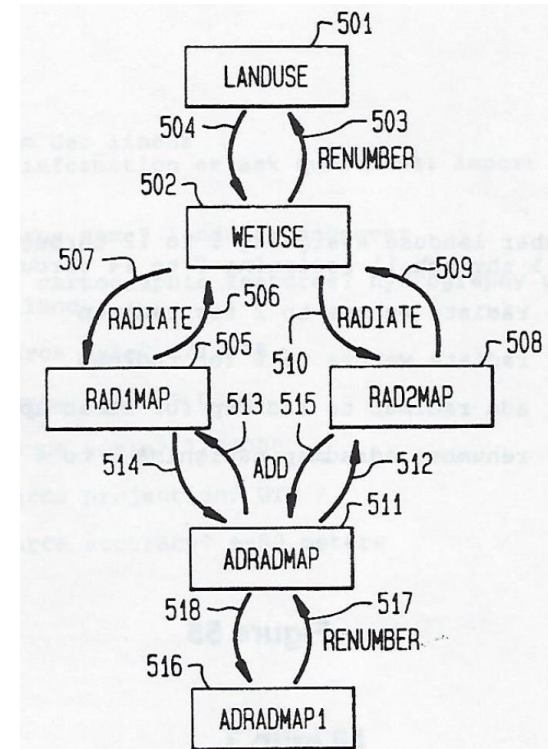


Querying metadata...

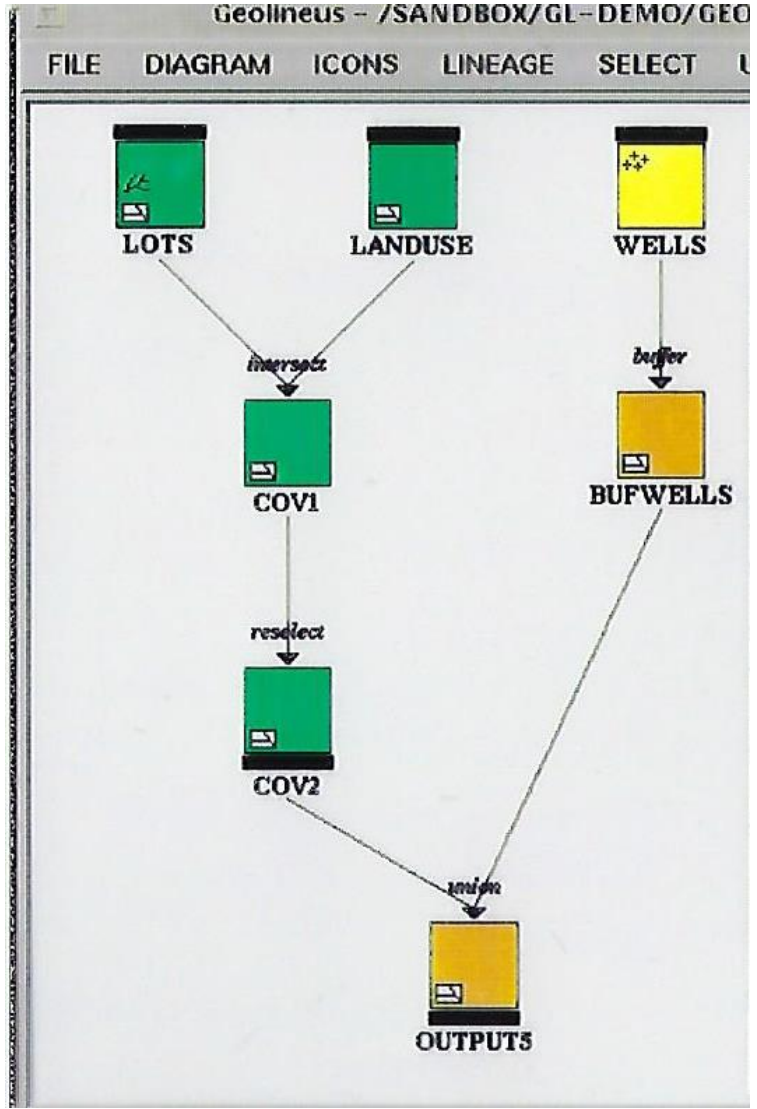
What are the final products of landuse
(ADRADMAP1 IS A PRODUCT MAP LAYER)



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.		GRID scalar variable.
	Derived layer. Layer was created as a result of an ARC/INFO command like BUFFER, INTERSECT or GRIDPOLY.		Coverage has been edited in ARCEDIT since the last CLEAN and BUILD.
	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 has been edited in ARCEDIT since the last CLEAN and BUILD and polygon topology needs rebuilding.
	Coverage containing point features. It has a point attribute table (PAT).		Coverage in which arc features have been rebuilt but polygon topology still needs rebuilding.
	Coverage containing arc features. It has an arc attribute table (AAT).		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.
	Coverage containing polygon features. It has a polygon attribute table (PAT).		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
	Coverage with both a point attribute table and an arc attribute table.		A 'dimmed' layer. This layer no longer exists. It has either been KILLED, or moves to a new location. Dimmed derived layers are recreated with the 'Recreate' option from the 'Update' menu.
	Coverage with both an arc attribute table and a polygon attribute table.		A dimmed GRID scalar. Icon was added to diagram with the 'Create from log' option so value is unknown
	Grid with integer cell values.		
	Grid with integer cell values, and a value attribute table (VAT)		
	Grid with floating point cell values.		

GUI design by Rupert Essinger

OK

Working with source and command metadata

The screenshot shows the Geolineus interface with a workflow diagram on the left. The workflow starts with 'LOTS', 'LANDUSE', and 'WELLS'. 'LOTS' and 'LANDUSE' are combined via an 'intersect' operation to create 'COV1'. 'COV1' is then processed by a 'reselect' operation to create 'COV2'. Finally, 'COV2' and 'WELLS' are combined via a 'union' operation to create 'OUTPUTS'. A 'Source Frame - LOTS' dialog box is open in the foreground, displaying metadata for the 'LOTS' source. A yellow callout box points to the 'DESCRIPTION' field.

Source Frame - LOTS

NAME: LOTS

DESCRIPTION

DATA QUALITY

SPATIAL EXTENT

MAP PROJECTION

DATUM

STATUS

POINT/VECTOR OBJECTS

CONTACT

ENTITY ATTRIBUTES

DESCRIPTION: 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

DATE: Fri 1-Apr-1994 14:00

DATE: Thu 15-Dec-1994 14:21

Buttons: OK, Import..., Cancel

This is where CIA source metadata would be added...

The screenshot shows the Geolineus interface with a workflow diagram on the left. The workflow starts with 'LOTS', 'LANDUSE', and 'WELLS'. 'LOTS' and 'LANDUSE' are combined via an 'intersect' operation to create 'COV1'. 'COV1' is then processed by a 'reselect' operation to create 'COV2'. 'WELLS' is processed by a 'buffer' operation to create 'BUFWELLS'. Finally, 'COV2' and 'BUFWELLS' are combined via a 'union' operation to create 'OUTPUTS'. A 'Command Frame - BUFWELLS' dialog box is open in the foreground, displaying metadata for the 'BUFWELLS' command. A yellow callout box points to the 'NOTE' field, and a red circle highlights the 'Ripple...' button.

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

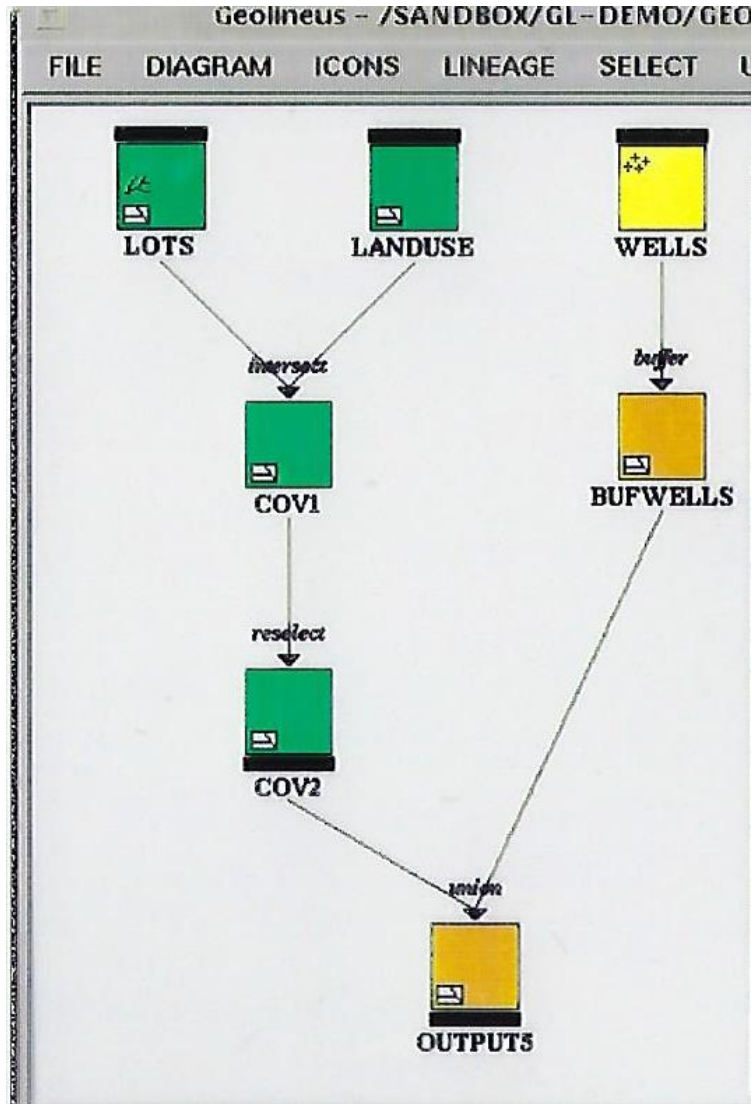
FIRST_CREATED: Sun 28-Apr-1991 16:33

LAST_RECREATED: Mon 29-Apr-1996 11:39

Buttons: OK, **Ripple...**, Cancel

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

Update propagation...



Geolineus - /SANDBOX/GL-DEMO/GEOLINEUS30/DEMO/DEMO3.LNG

FILE DIAGRAM ICONS LINEAGE SELECT UPDATE DELETE HELP

LOTS LANDUSE WELLS

intersect

COV1

buffer

BUFWELLS

resselect

COV2

union

OUTPUTS

Commands to update data

buffer wells bufwells # # 120 # point

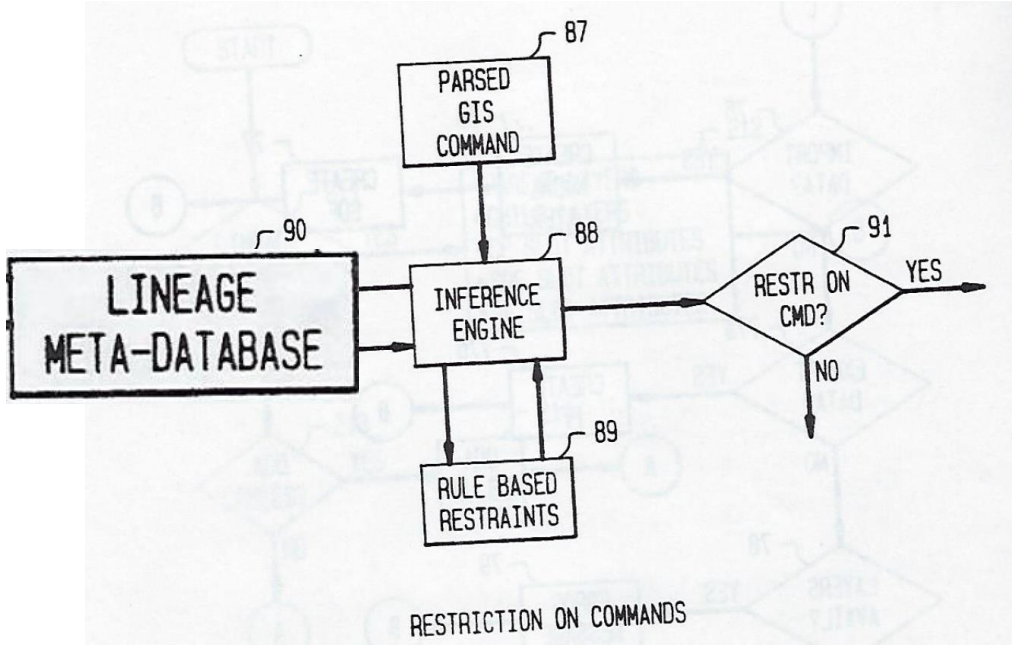
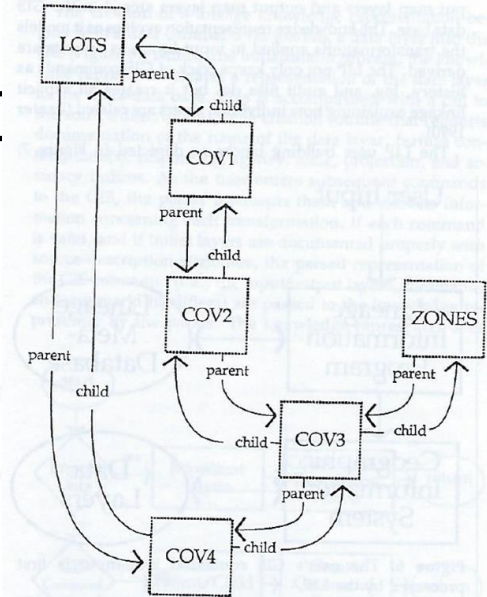
OK

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

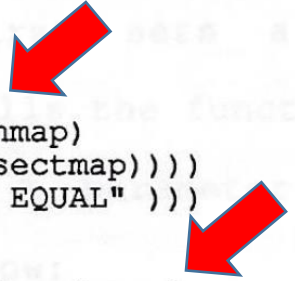
```
Killed bufwells with the ARC option
Arc: buffer wells bufwells # # 120 # point
Buffering ...
Sorting...
Intersecting...
Assembling polygons...
Creating new labels...
Finding inside polygons...
Dissolving...
Creating bufwells.PAT...
Arc: union bufwells cov2 output5
Unioning bufwells with cov2 to create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

Detailed description: This screenshot shows the Geolineus workspace with the same data flow as the diagram. A dialog box titled 'Commands to update data' is open, displaying the command 'buffer wells bufwells # # 120 # point' and an 'OK' button. At the bottom, the ARC/INFO command window shows the execution of these commands, including '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...', and 'Intersecting...'.

Data source metadata based integrity constraint

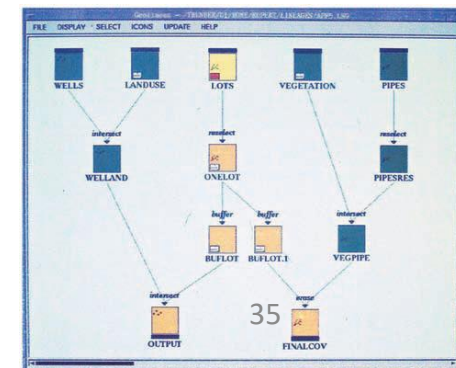
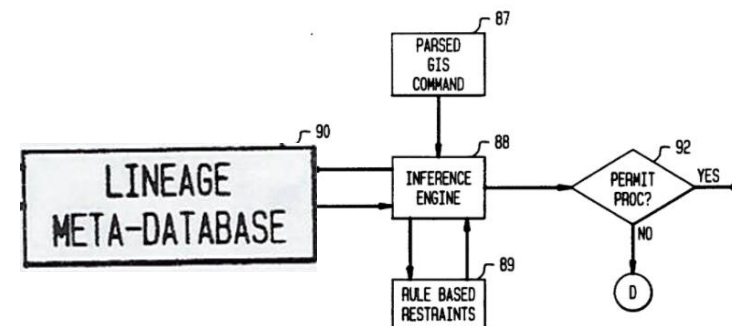



```
(setq intersect_rules
  '((rule intersect1
    (if (not (equal (scale inmap)
                    (scale intersectmap))))
    (then ("INPUT SCALES NOT EQUAL" )))
    (rule intersect2
    (if (not (equal (projection inmap)
                    (projection intersectmap))))
    (then ("INPUT PROJECTIONS NOT EQUAL"
           ("Reproject one of the maps.")) )))
```



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

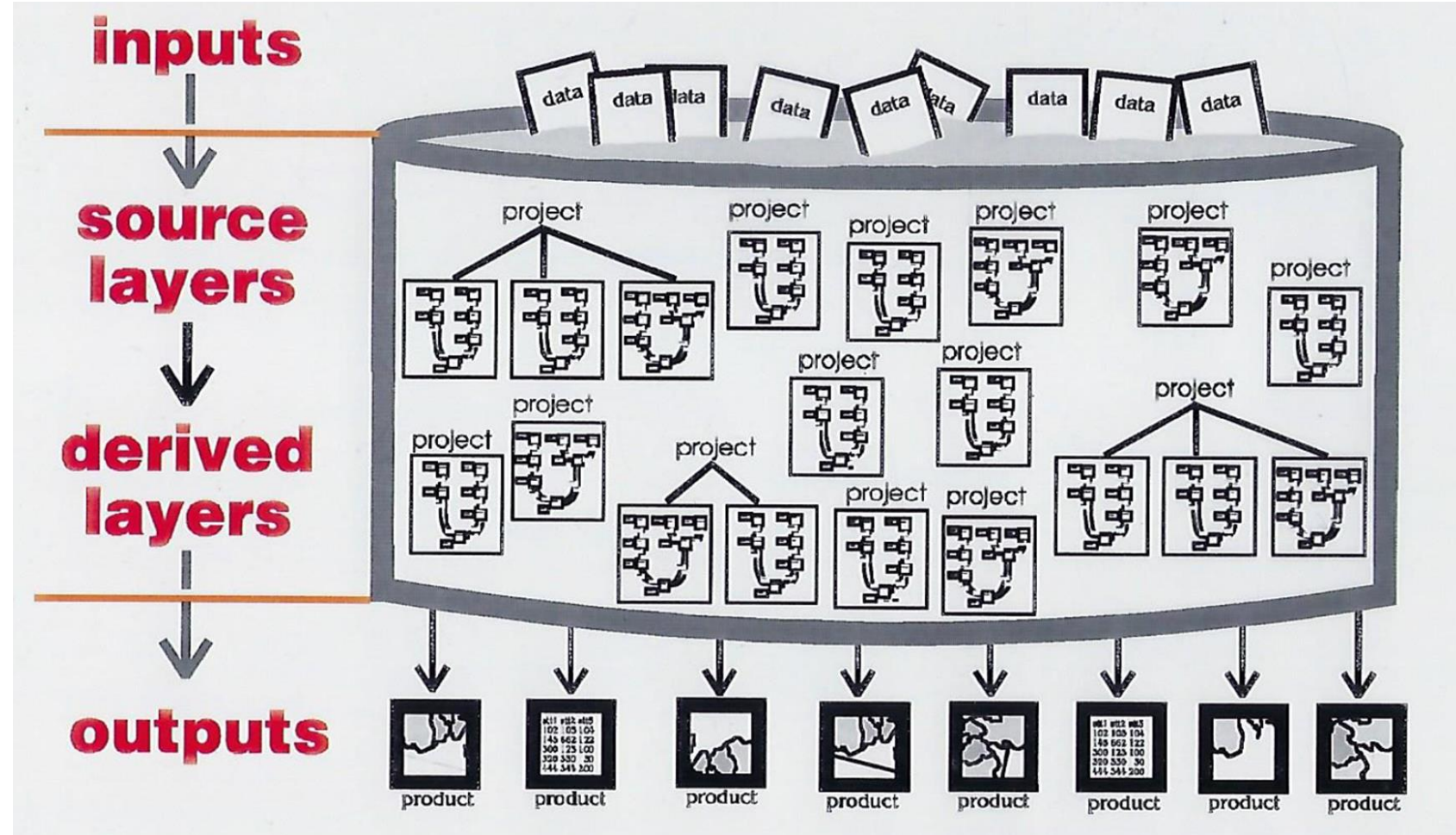
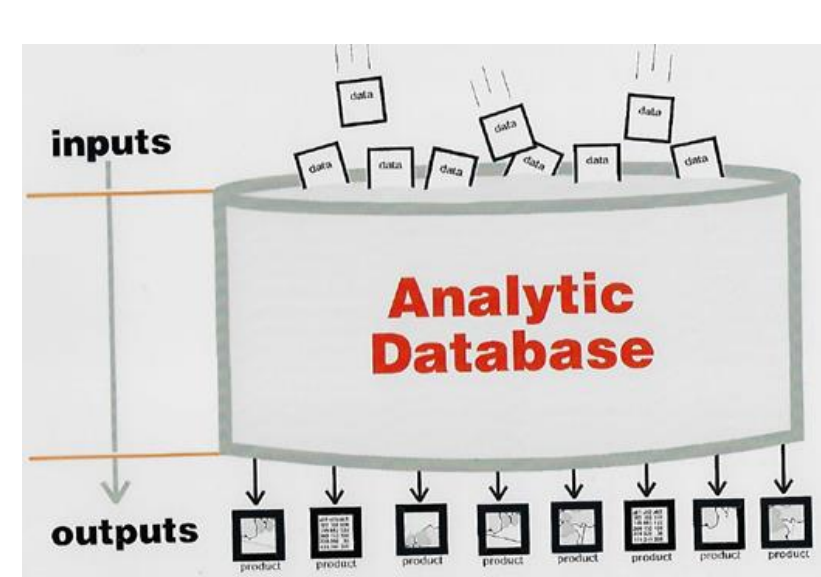


at Southern California Edison

Focus of the audit:

1. Documentation and understanding of 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

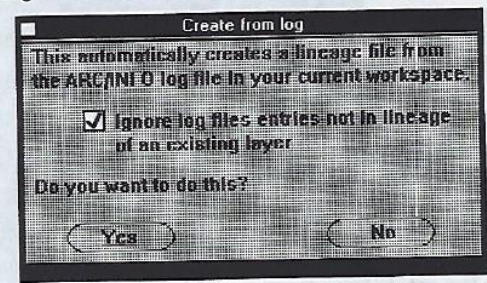
- What is Geolineus? 3
- What does a lineage diagram show? 4
- How does Geolineus store metadata? 9
- Working with Geolineus 11
- Geolineus demo 13
- Creating frame templates 19
- Creating a new lineage diagram 22
- Documenting source data 24
- Documenting derived data 26
- Documenting product data 29
- 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 (↓).

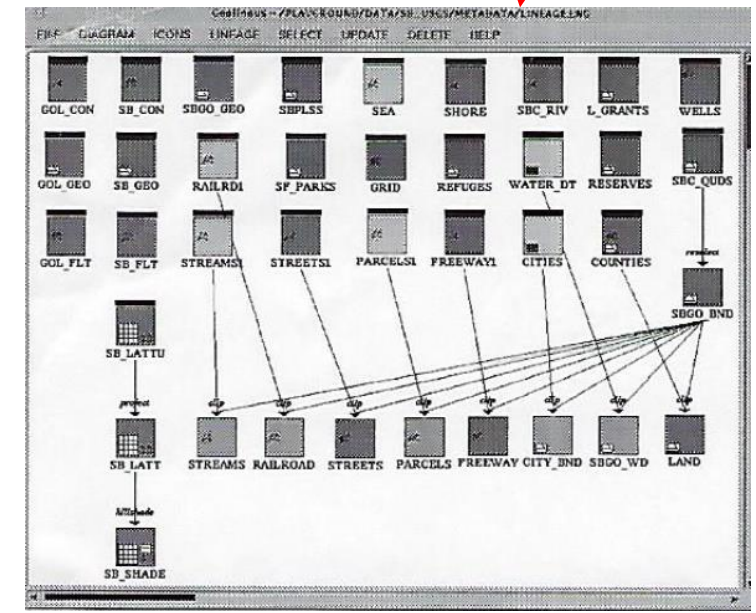


The check option enables you to choose whether or not the diagram that "Create from log" will create will include lineage for layers that no longer exist. Normally, Geolineus will ignore no lines in the log file that do not contribute to the lineage of an existing layer. This results in a lineage diagram 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 for example to create a diagram from a log file for which the data is unavailable.

Log Files

198923021442	1	3	OARCPLT
198923021442	0	10	OBUILD NISLAND POLY
198923021442	0	1	OEXTERNAL NISLAND
198923021503	20	44	OARCPLT
198923021505	0	3	OPOLYGRID NISLAND
198923021512	2	15	Opolygrid nisland
198923021514	1	24	Ogridpoly nisland.svf nigrd 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 nigrd
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



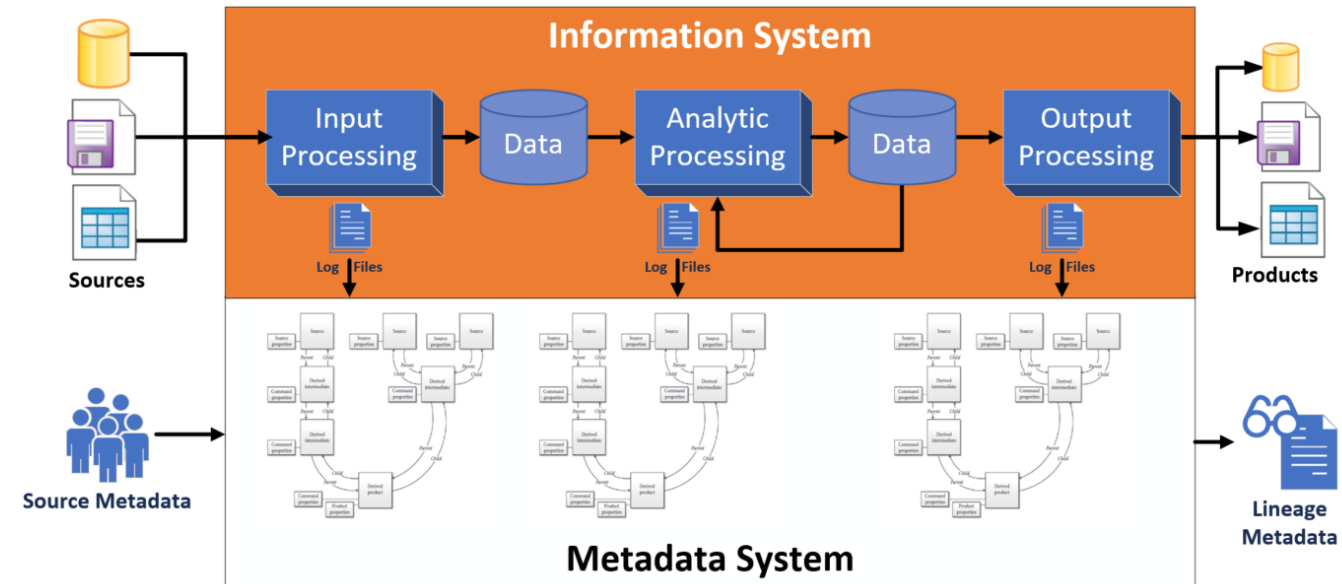
Lineage metadata enabled audit of data and processing



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

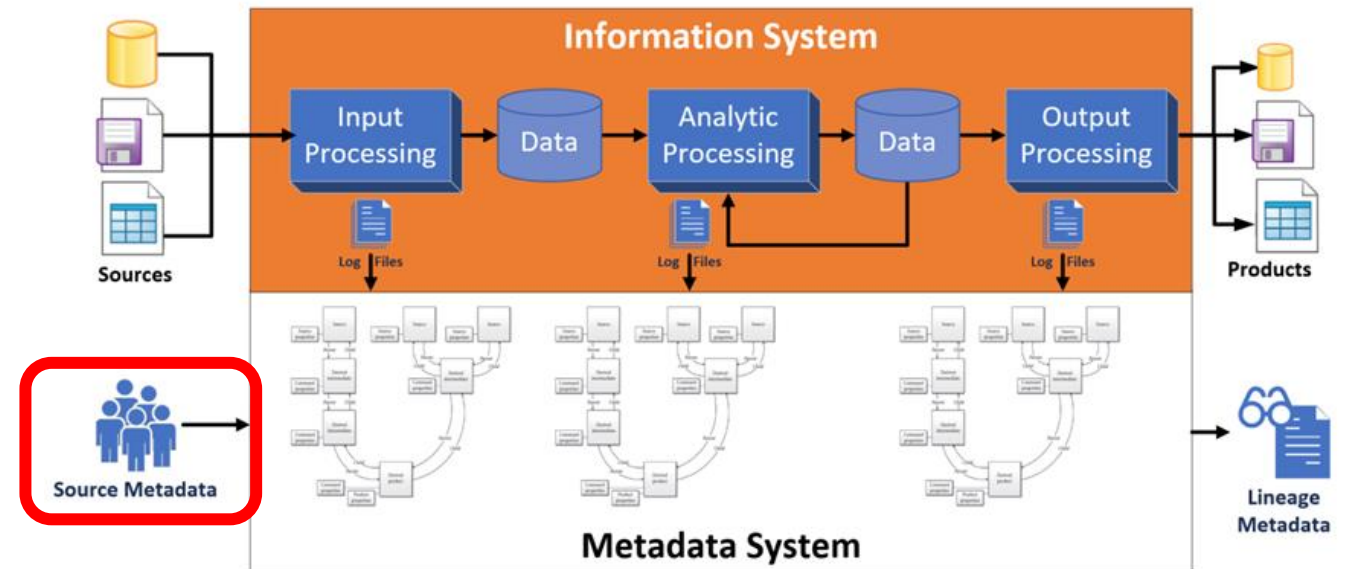
Lineage metadata enabled audit of data and processing



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



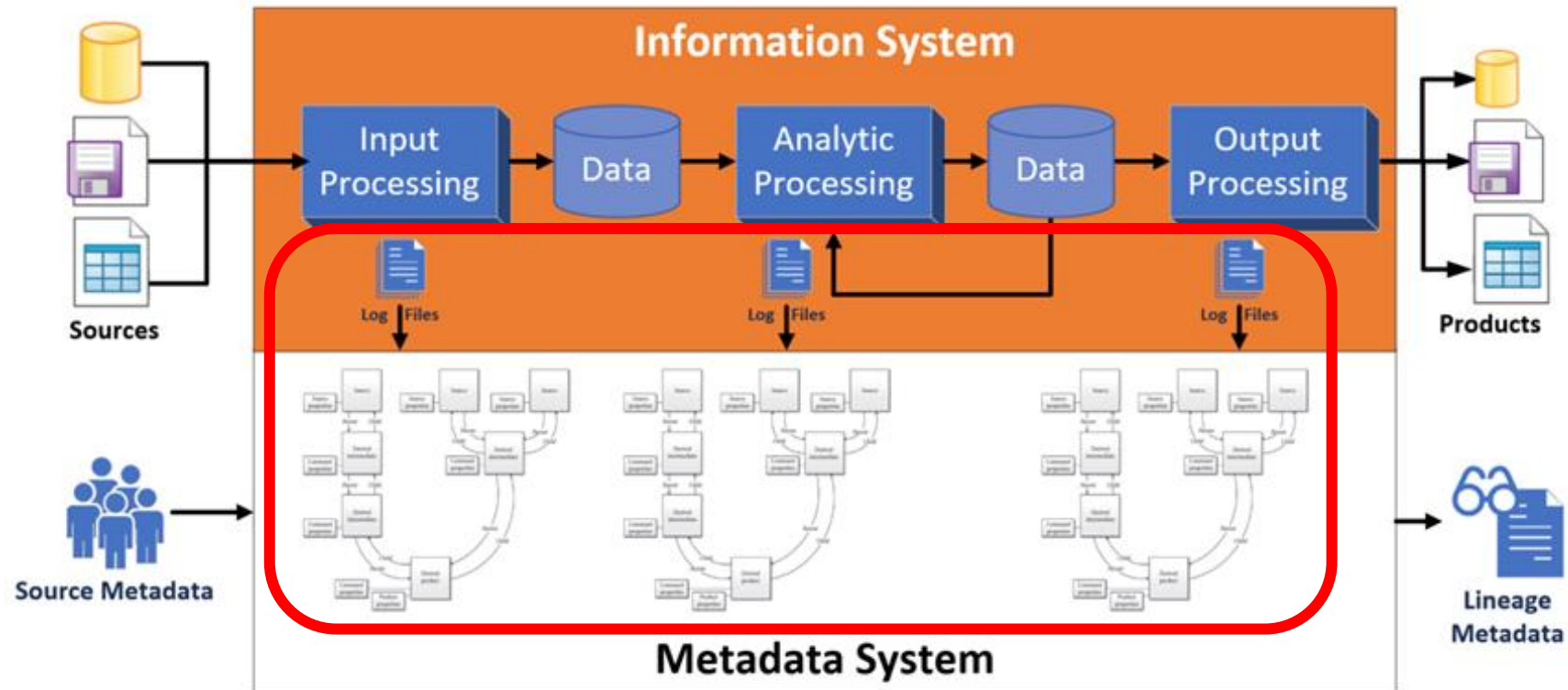
Metadata enabled audit of data and processing



at Southern California Edison

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

Reverse engineer lineage metadata from the log files



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:

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

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

and,

$X_{source} = (A_{source\ features}, A_{source\ date}, \dots, A_{source\ accuracy}) \subset A_{source}$
 $l_{source'} \equiv l_{source''} \text{ iff } \forall A_{source\ k} \in X_{source} \wedge a_{source'\ k} = a_{source''\ k}$

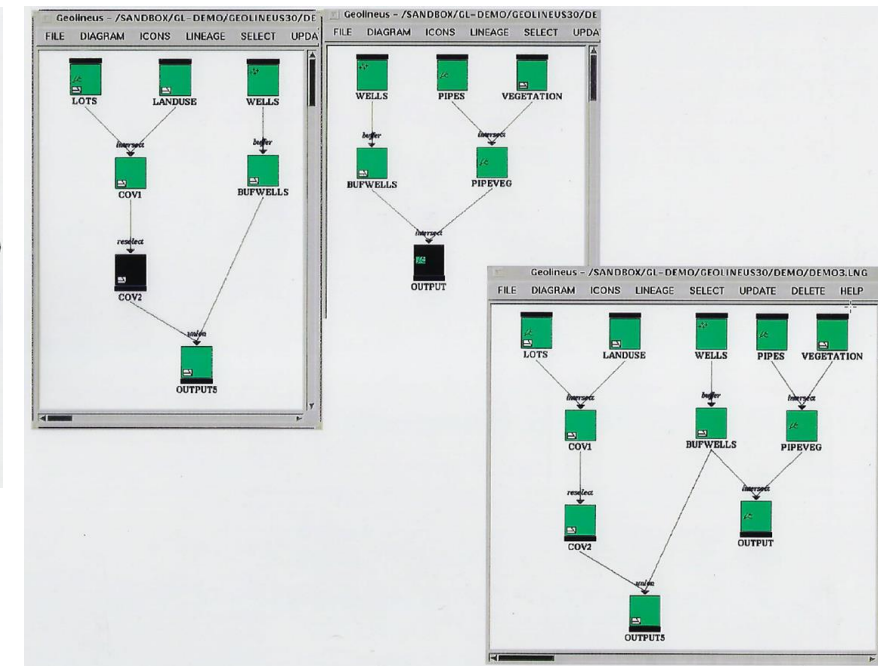
Source equivalence testing

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

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

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

Derived equivalence testing



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

Findings:

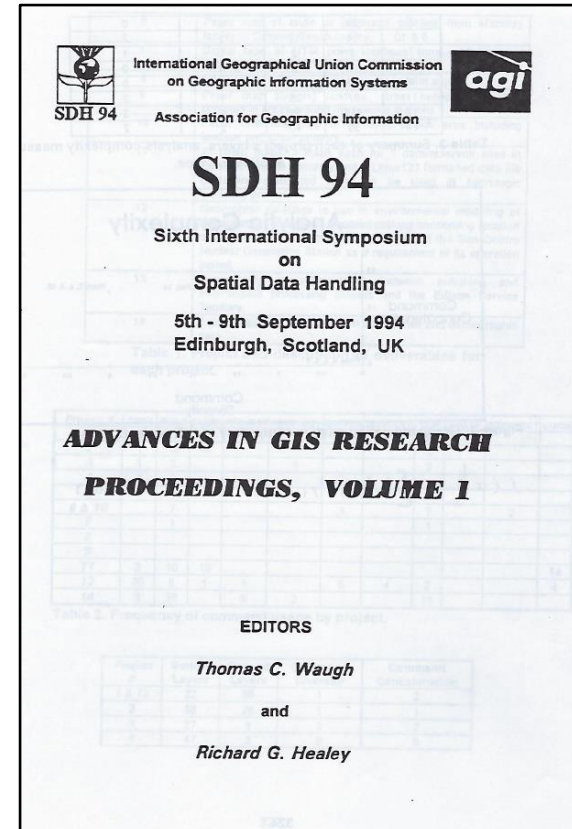
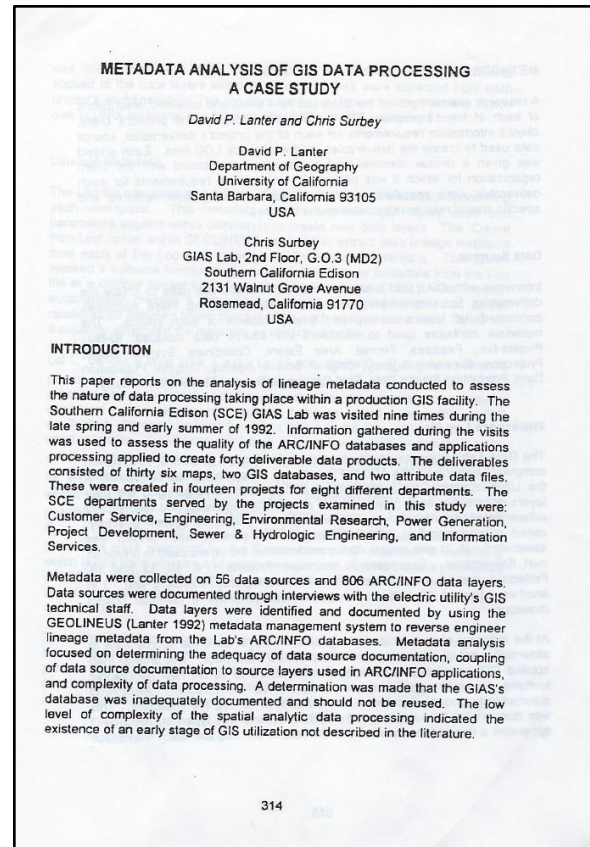
2. Lack of naming conventions for identifying primary data source files and source datasets once they were imported into the Information System

- 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 time to visually inspect and compare the actual data – it was not clear what, if any, significant differences existed among the versions

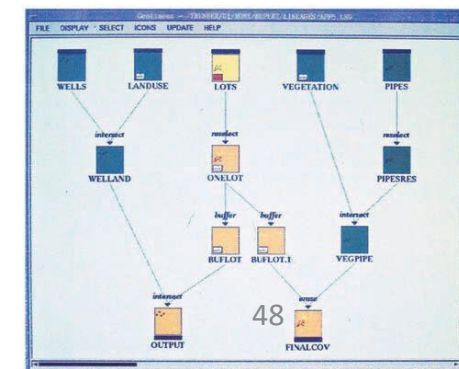
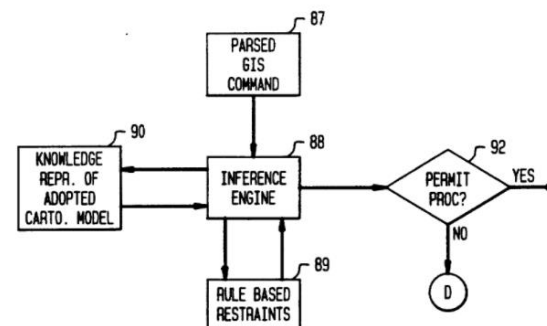
Recommendation:

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



Conclusion: 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

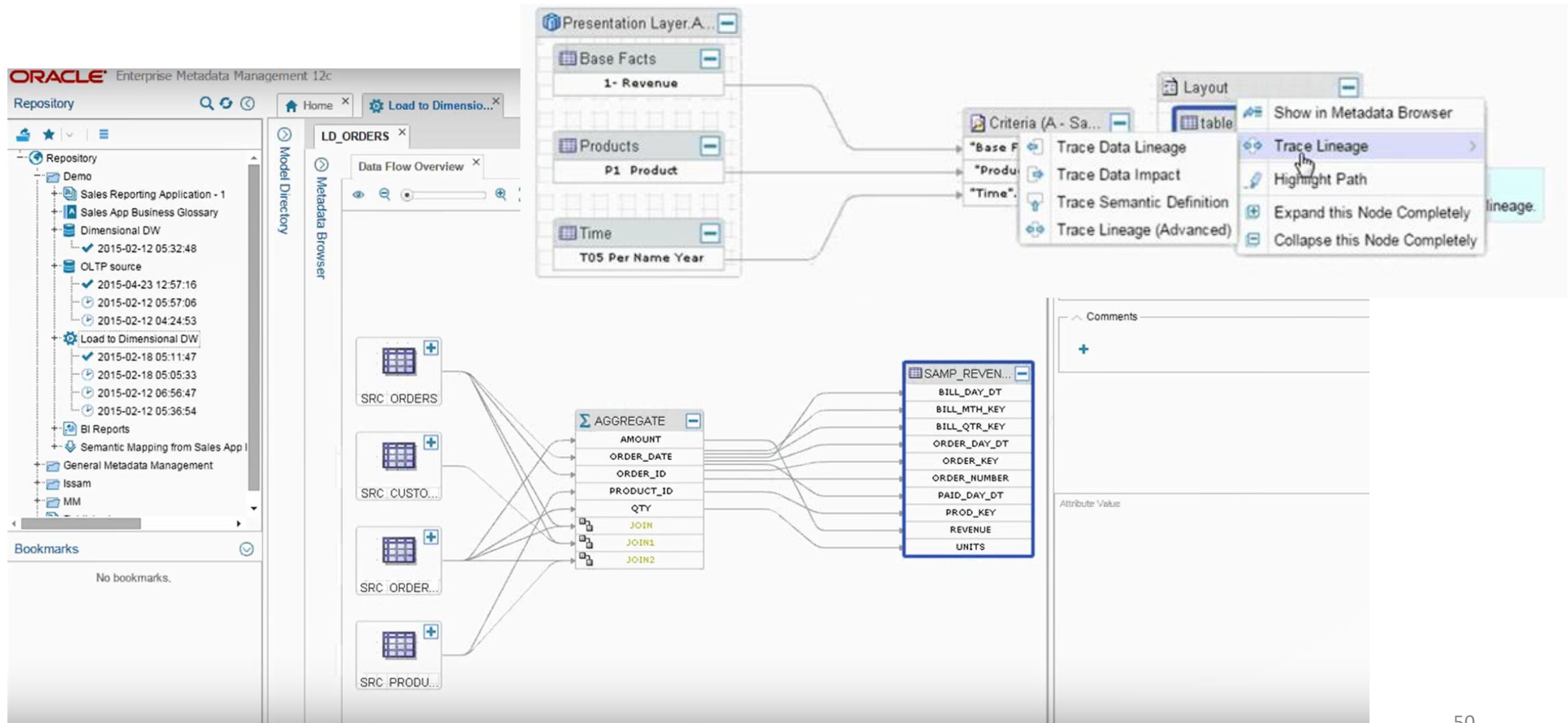


Conclusion:

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 ...

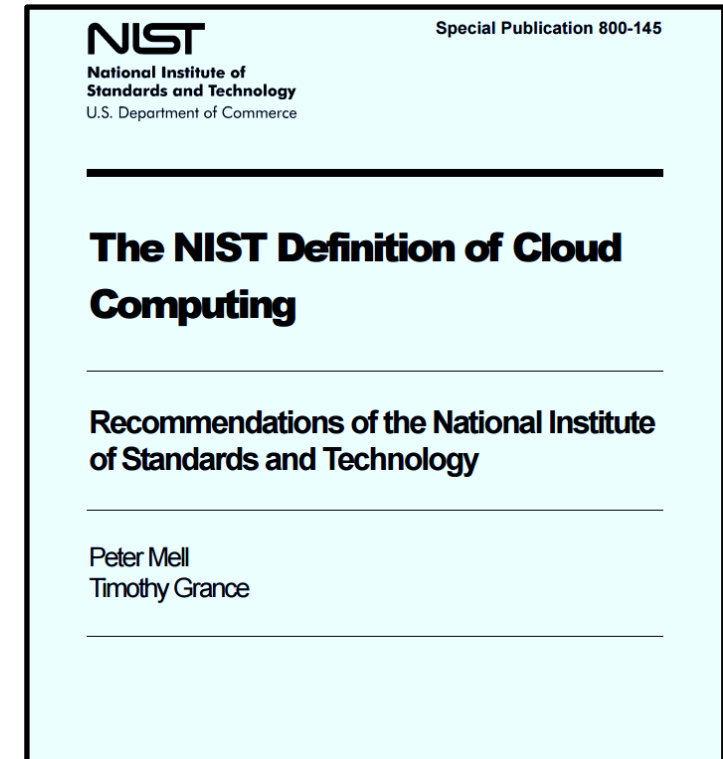


Agenda

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

Cloud computing

Cloud computing enables ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction



Essential Characteristics of Cloud Computing

- 1. On-demand self-service**
- 2. Broad network access**
- 3. Resource pooling**
- 4. Rapid elasticity**
- 5. Measured service**

Which Service Model(s) of cloud computing is your project's information system providing to your end users?

Table 8-1 Service Layers Represented in this SSP

Service Provider Architecture Layers		
<input type="checkbox"/>	Software as a Service (SaaS)	Major Application
<input type="checkbox"/>	Platform as a Service (PaaS)	Major Application
<input type="checkbox"/>	Infrastructure as a Service (IaaS)	General Support System
<input type="checkbox"/>	Other	Explain: Click here to enter text.

3 Service Models of Cloud Computing

Infrastructure as a Service (IaaS)

Provides processing, storage, networks, and other fundamental computing resources

Consumer is able to deploy and run arbitrary software, which can include operating systems and applications

- The consumer does not manage or control the underlying cloud infrastructure,
 - but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls)

3 Service Models of Cloud Computing

Platform as a Service (PaaS)

Consumer is provided capability to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider

- The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage,
 - but has control over the deployed applications and possibly configuration settings for the application-hosting environment

3 Service Models of Cloud Computing

Software as a Service (SaaS)

The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure

- Accessible from various client devices through either a thin client interface, such as a web browser or a program interface
- The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user specific application configuration settings

Which cloud deployment model is your project's information system based on?

8.2 CLOUD DEPLOYMENT MODELS

Information systems are made up of different deployment models. The deployment models of the Information System Abbreviation that are defined in this SSP and are not leveraged by any other FedRAMP Authorizations, are indicated in Table 8-2 Cloud Deployment Model Represented in this SSP that follows.

Instruction: Check deployment model that applies.

Delete this and all other instructions from your final version of this document.

Table 8-2 Cloud Deployment Model Represented in this SSP

Service Provider Cloud Deployment Model		
<input type="checkbox"/>	Public	Cloud services and infrastructure supporting multiple organizations and agency clients
<input type="checkbox"/>	Private	Cloud services and infrastructure dedicated to a specific organization/agency and no other clients
<input type="checkbox"/>	Government Only Community	Cloud services and infrastructure shared by several organizations/agencies with same policy and compliance considerations
<input type="checkbox"/>	Hybrid	Explain: (e.g., cloud services and infrastructure that provides private cloud for secured applications and data where required and public cloud for other applications and data) Click here to enter text.

4 Deployment Models of Cloud Computing

Public cloud

The cloud infrastructure is provisioned for open use by the general public

- It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider

4 Deployment Models of Cloud Computing

Private cloud

The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units)

- It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

4 Deployment Models of Cloud Computing

Community cloud

Provisioned for use by a specific community of consumers from organizations with shared concerns

- It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises

4 Deployment Models of Cloud Computing

Hybrid cloud

A composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities

- ...but are bound together by standardized or proprietary technology that enables data and application portability

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

Control Originatoin

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

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): <input type="checkbox"/> Implemented <input type="checkbox"/> Partially implemented <input type="checkbox"/> Planned <input type="checkbox"/> Alternative implementation <input type="checkbox"/> Not applicable	
Control Origination (check all that apply): <input type="checkbox"/> Service Provider Corporate <input type="checkbox"/> Service Provider System Specific <input type="checkbox"/> Service Provider Hybrid (Corporate and System Specific) <input type="checkbox"/> Configured by Customer (Customer System Specific) <input type="checkbox"/> Provided by Customer (Customer System Specific) <input type="checkbox"/> Shared (Service Provider and Customer Responsibility) <input type="checkbox"/> Inherited from pre-existing <u>FedRAMP</u> 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**