MIS 4596

Data Privacy

Class 5

Agenda

- Online privacy
- Privacy and data protection by design

  …with data provenance and lineage metadata

ONLINE PRIVACY:
HOW DID WE GET HERE?

California Consumer Privacy Act (CCPA, 2018)
&
California Privacy Rights Act (CPRA, 2020)

California voters approve Prop. 24, ushering in new rules for online privacy

Concordia University coronavirus 'outbreak' attributed to more than 50 'false positives'

Are L.A. County's new COVID restrictions rea necessary? We talk to the experts

Coronavirus infections are higher than ever, COVID-19 deaths are not. Why?

State Data Privacy Laws in Effect

- **California Consumer Privacy Act (CCPA) – Effective date: 1/1/2020**
- California Privacy Rights Act (CPRA) – Effective date: 1/1/2023
- Colorado Privacy Act (CPA) – Effective date: 1/1/2023
- Virginia Consumer Data Protection Act – Effective date: 1/1/2023

US State Privacy Legislation Tracker 2024

Statute/bill in legislative process
- Introduced
- In committee
- In cross chamber
- In cross committee
- Passed
- Signed
- Inactive bills
- No comprehensive bills introduced
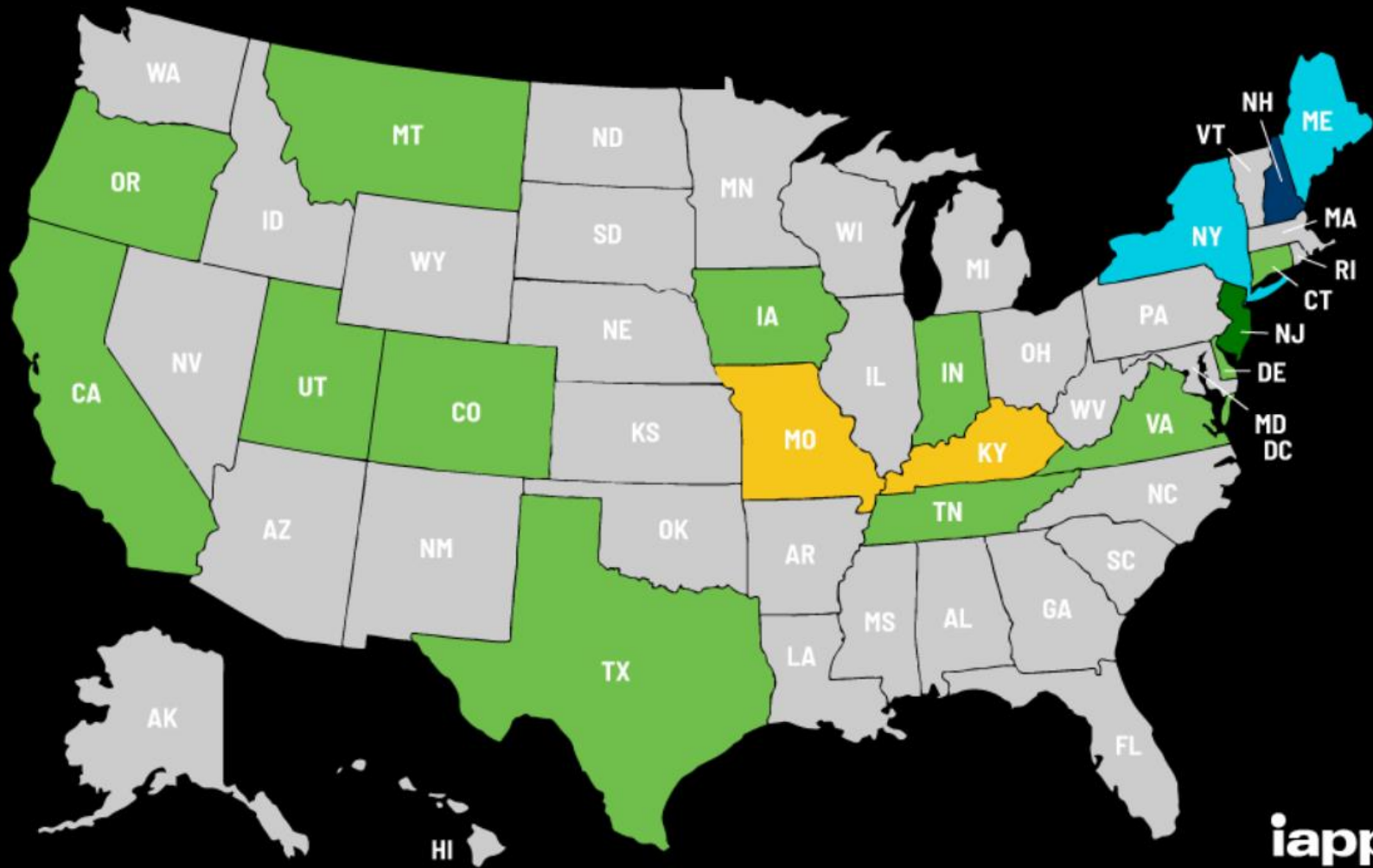
Last updated 9 Jan. 2024

iapp

https://iapp.org/media/images/resource_center/State_Comp_Privacy_Law_Map.png
https://iapp.org/resources/article/us-state-privacy-legislation-tracker/

Hilton Hotels fined for credit card data breaches

🕐 1 November 2017

f 💬 🐦 ✉ ⌁ Share

Hilton Hotels fined for credit card data breaches

1 November 2017

Share

Hilton's $700,000 fine for data breach impacting 350,000 customers

Hilton Hotels fined for credit card data breaches

1 November 2017

Share

Under the European Union's General Data Protection Regulation (GDPR) the fine would have been 4% of Hilton's global revenue $420,000,000

GDPR requires _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

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…
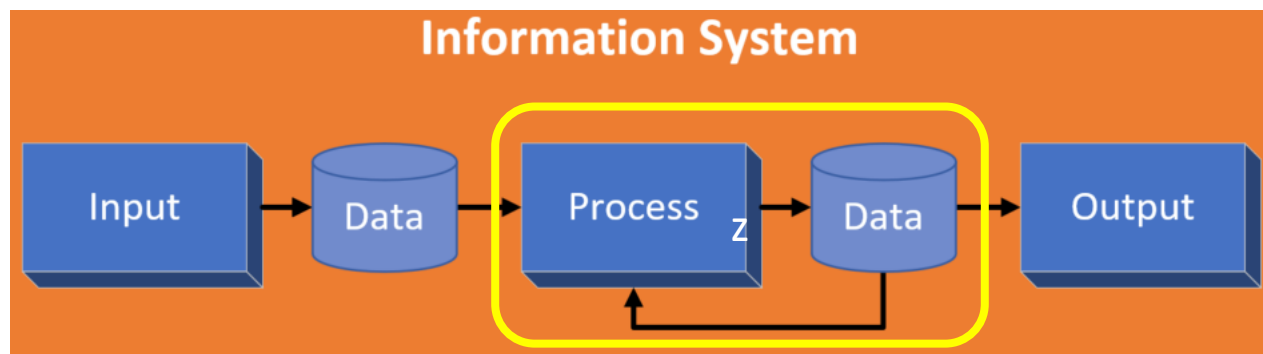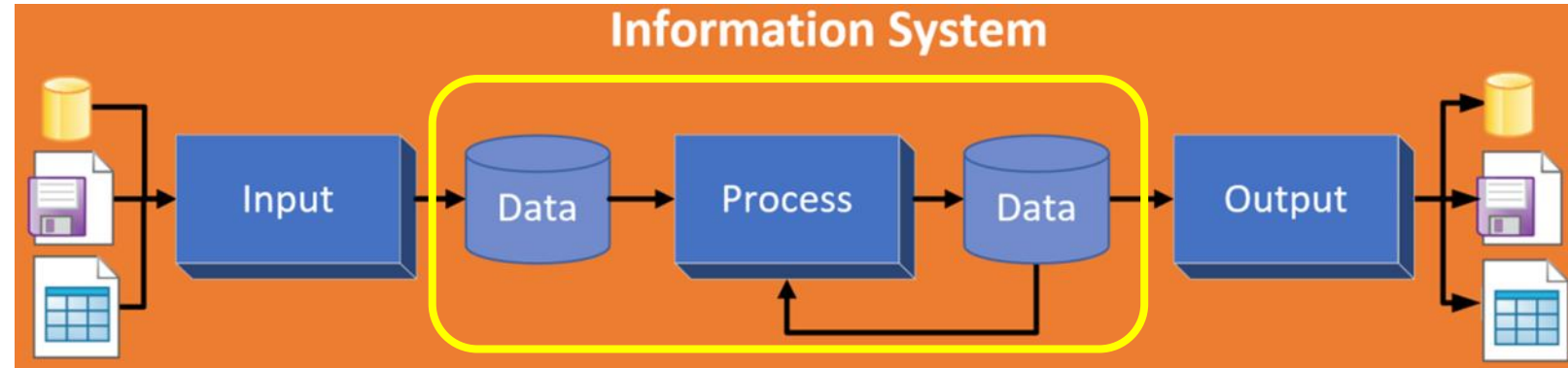

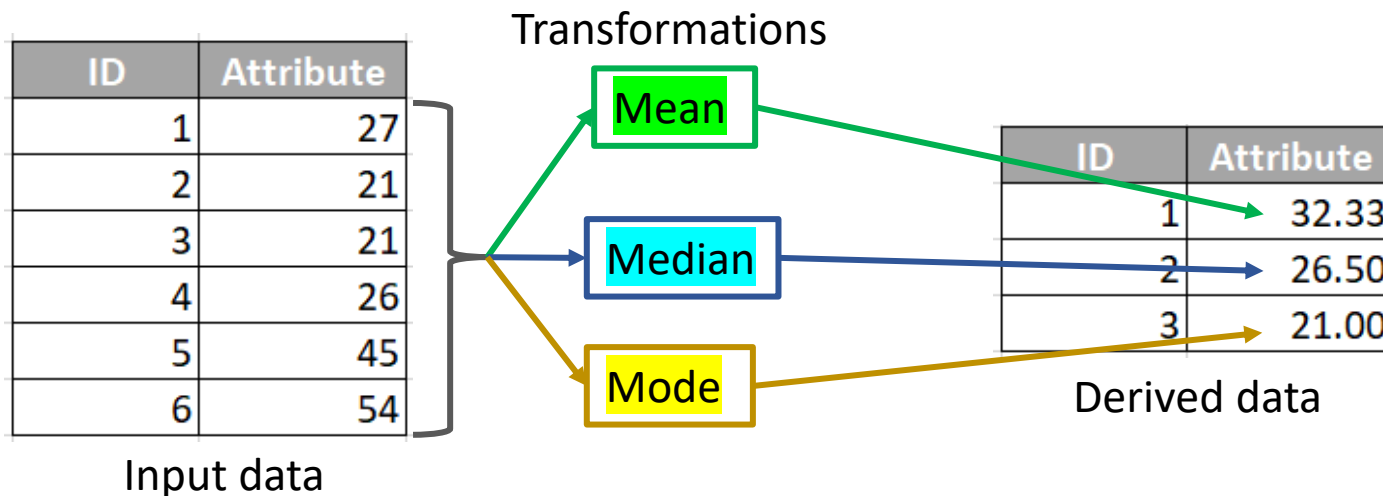Regardless of application, there is reliance on data processing workflows to produce and use information
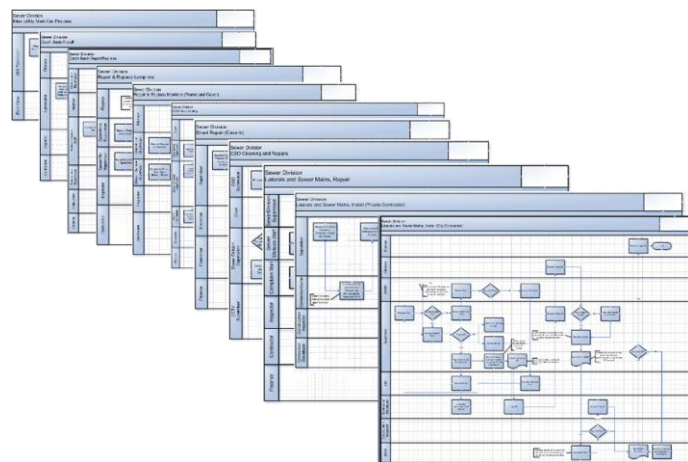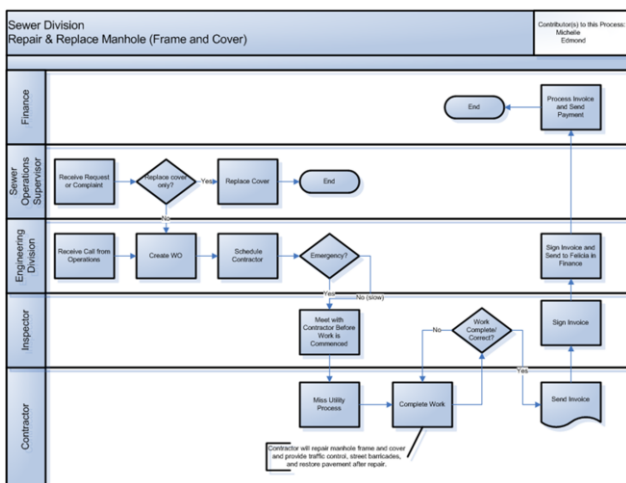
Data processing often transforms existing data into new data, which is a double-edged sword…

➢ *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…
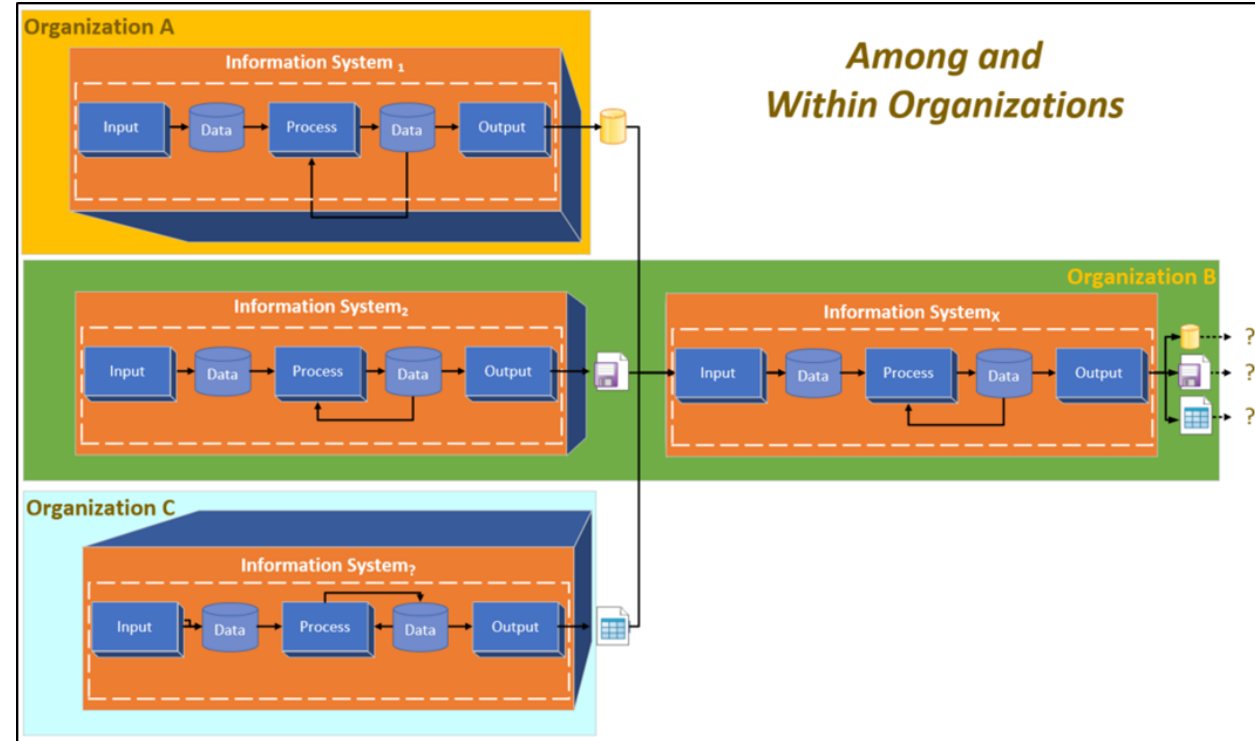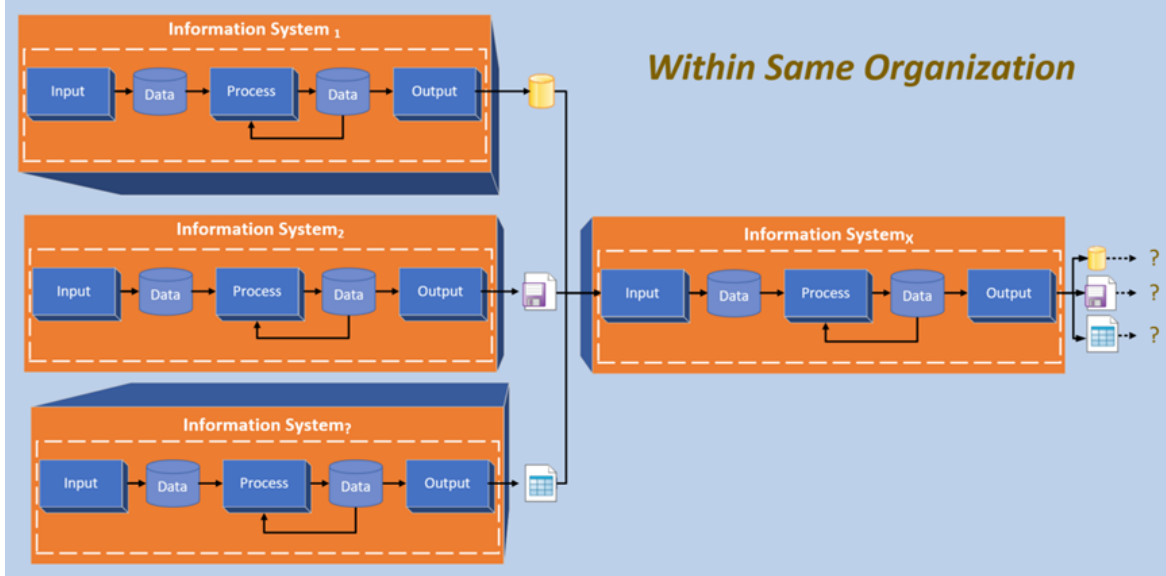
Provenance

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

Traditional Provenance

```
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
- *This 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/

W3C = World Wide Web Consortium

Provenance and data lineage

ELLIOTT GENEALOGY CHART

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

- Data provenance suggests <u>process history</u>
- Data lineage implies a kind of genealogy or data pedigree record relative to both:
  1. <u>Sources of data</u>
     +
  2. <u>Processing applied to the sources to produce an information product</u>
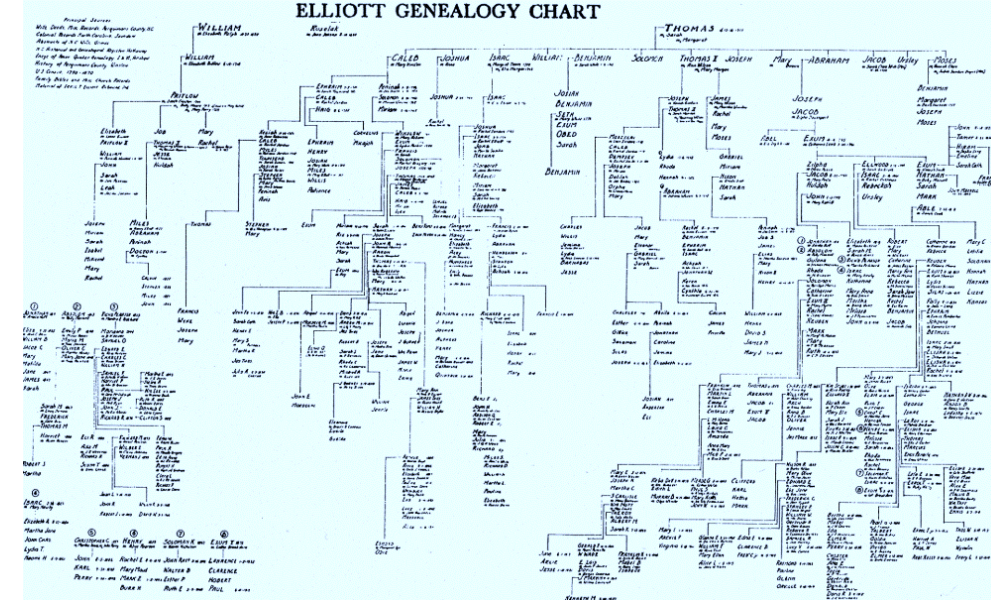
Data lineage metadata can aid understanding and establish trust of data…

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 —   Geoinformation —
Métadonnées —   Metadaten —
Partie 1: Principes de base   Teil 1: Grundsätze
(ISO 19115-1:2014)   (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.

cen

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
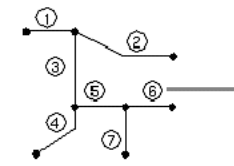
- Also organize their data in
  - Data base management systems
  - File systems
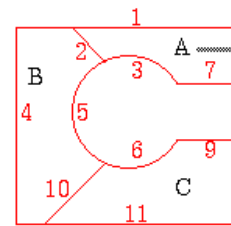
Polygon Attribute Table

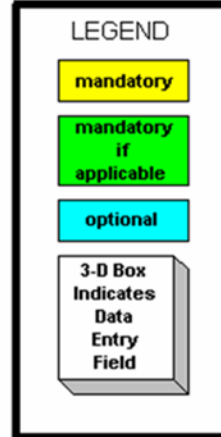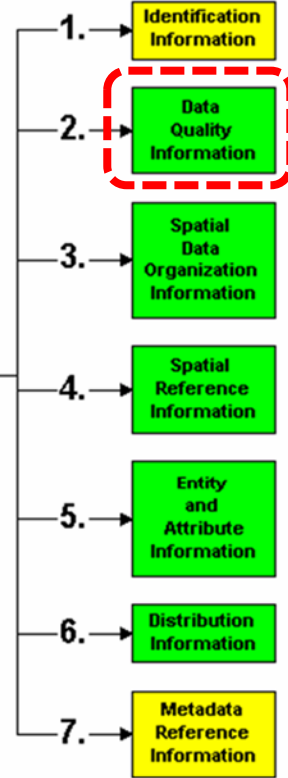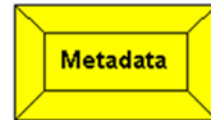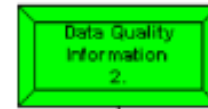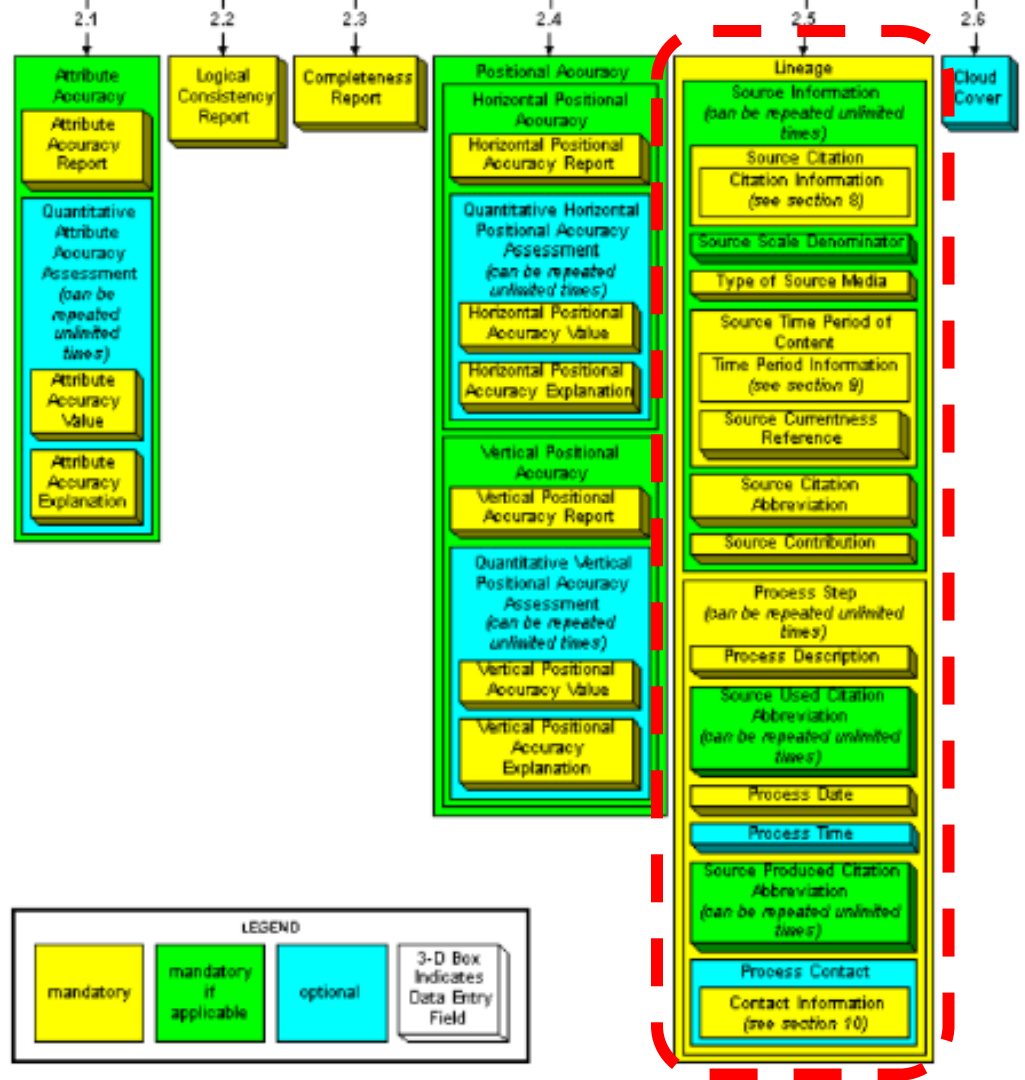| Polygon	Area	Parcel Number	Land Use
A	12,001	11-115-001	R 1
B	15,775	11-115-002	R 1
C	19,136	11-115-003	R 3

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

Content Standard for Digital Geospatial Metadata

FGDC-STD-001-1998

National Spatial Data Infrastructure

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

CSDGM Version 2 - 1998
(FGDC-STD-001 June 1998)

Data Quality Information

Data Quality Information 2.

- 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

1ˢᵗ application for tracking the lineage of data throughout their processing in information systems

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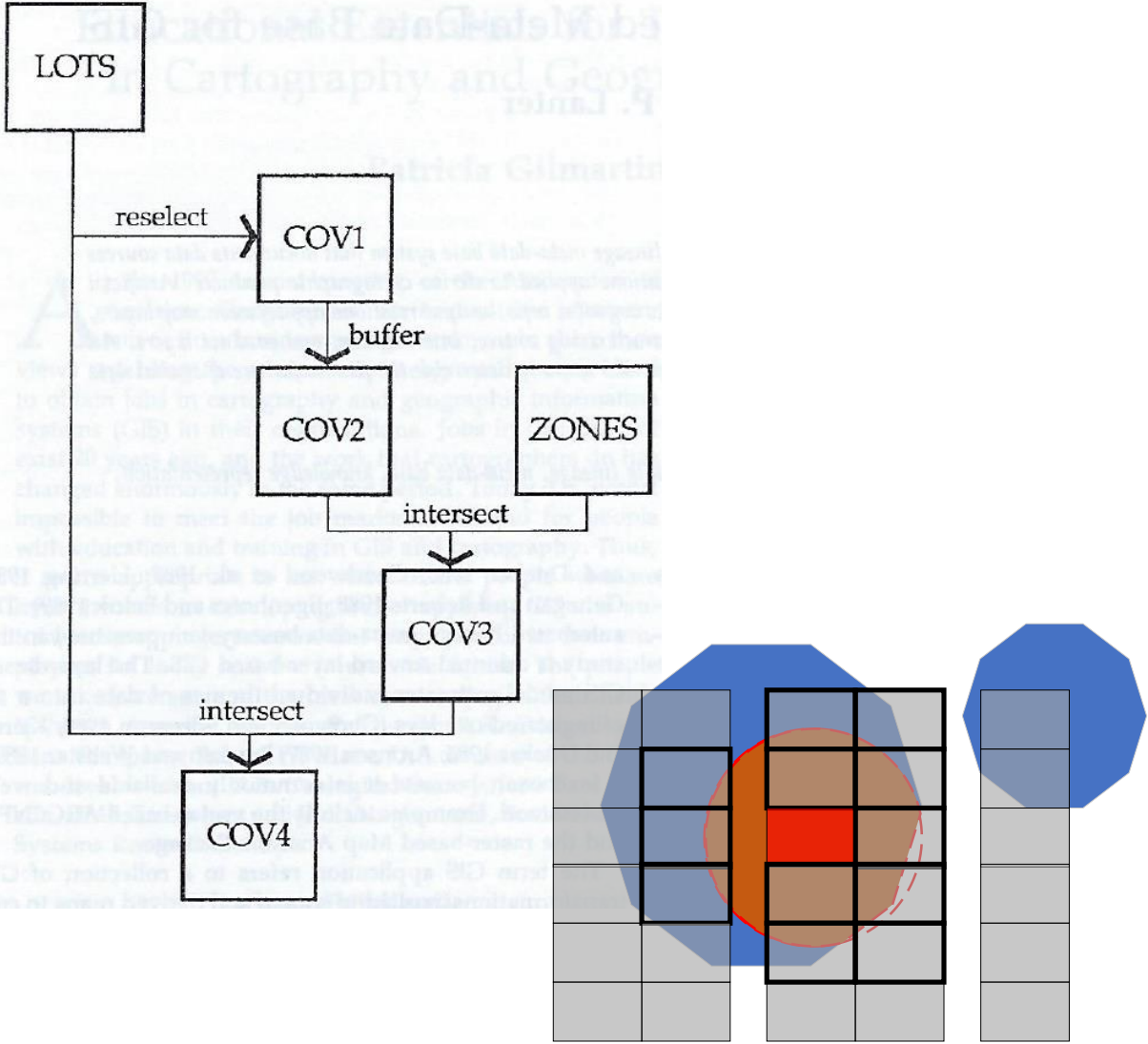
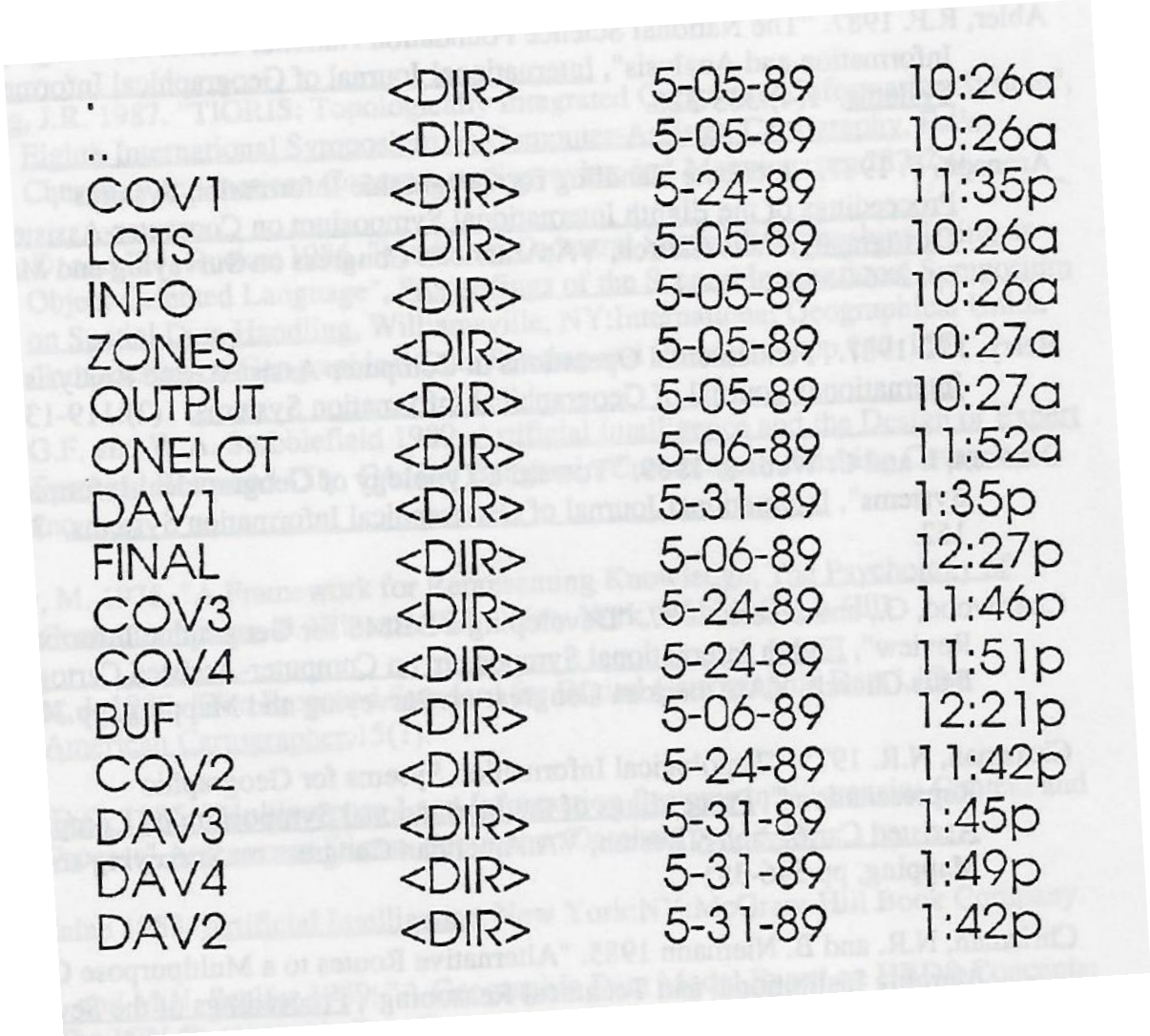Information processing steps in the head of the user as he transformed the LOTS and ZONES datasets to derive COV4…

Datasets stored on the computer after data processing analysis was over…

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

29

Datasets stored on the computer after data processing analysis was over…

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
```

After the work on the analysis was completed, it was difficult to remember which dataset was goal (i.e. end result)

How can I program the computer to help me remember what I knew about the data I loaded and processed on my computer?
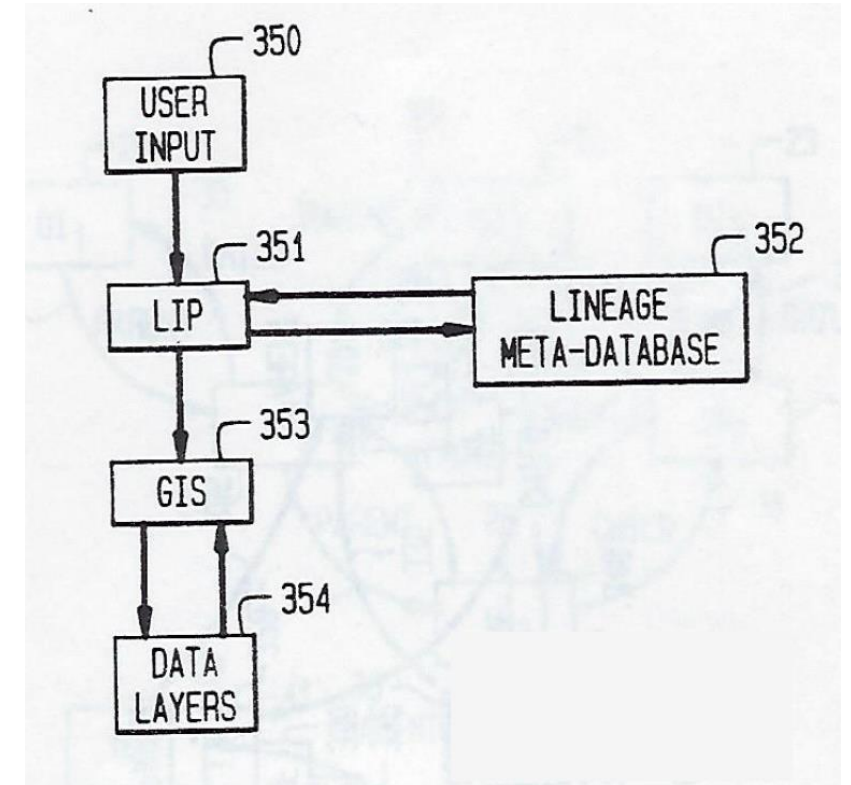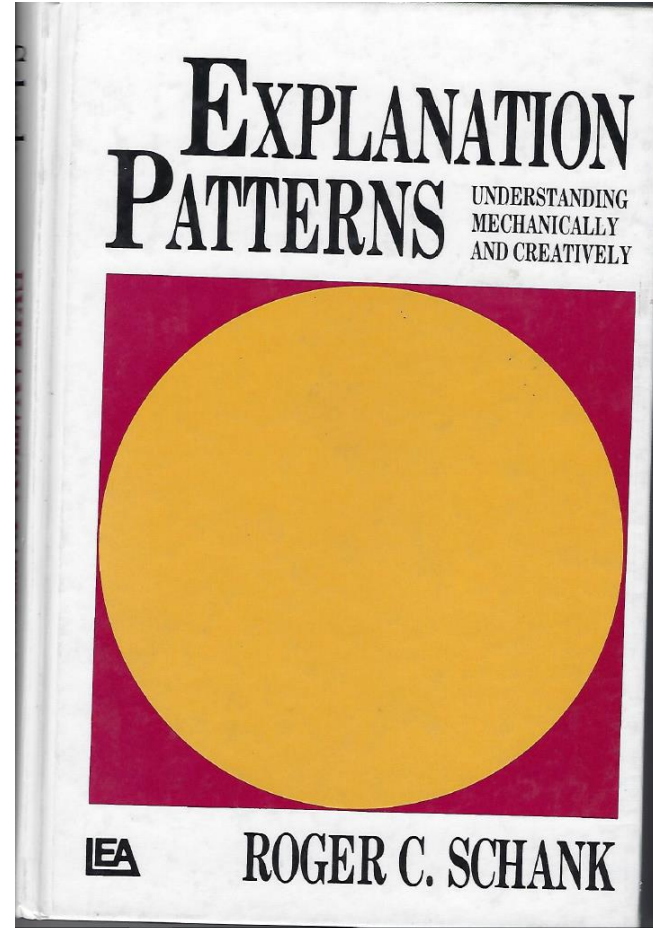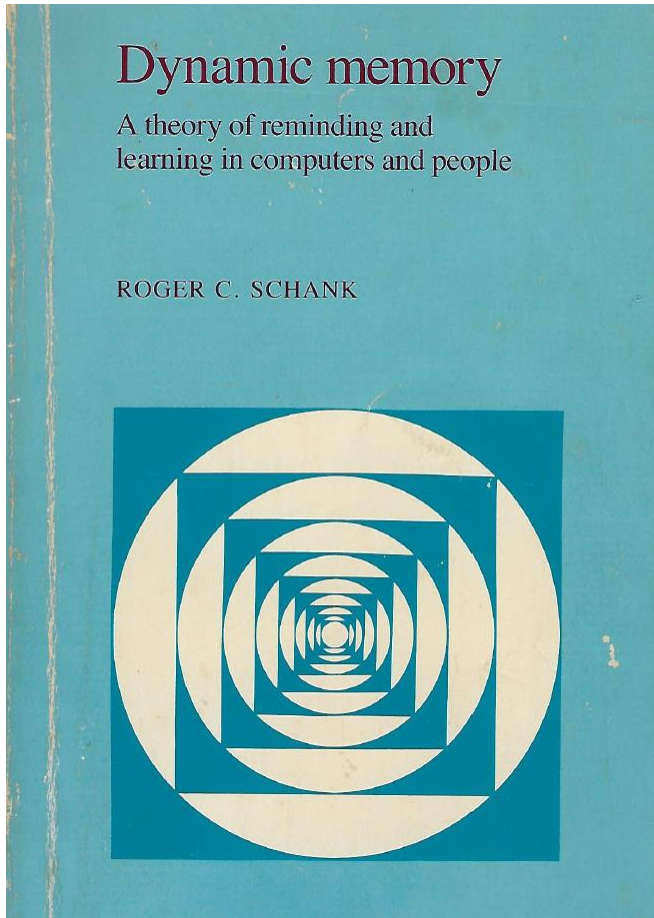
LIP = Lineage Information Processor

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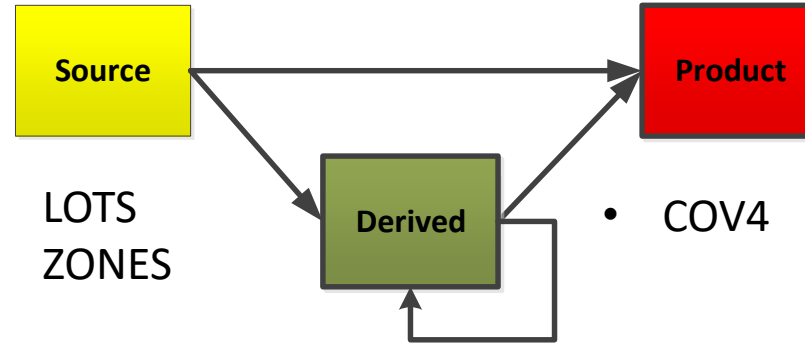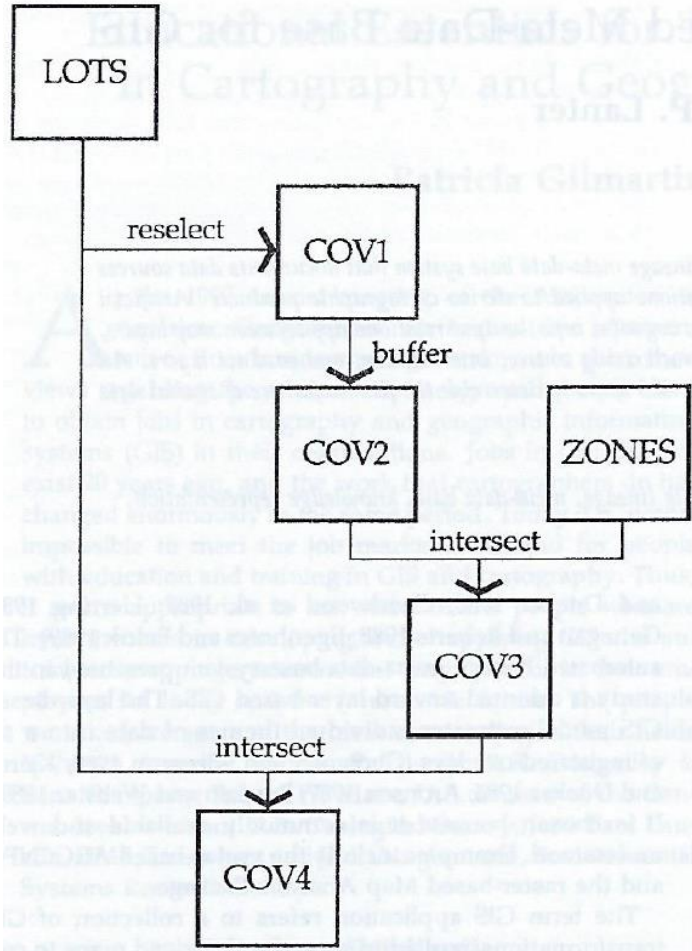
- LOTS
- ZONES

- COV1
- COV2
- COV3

- COV4
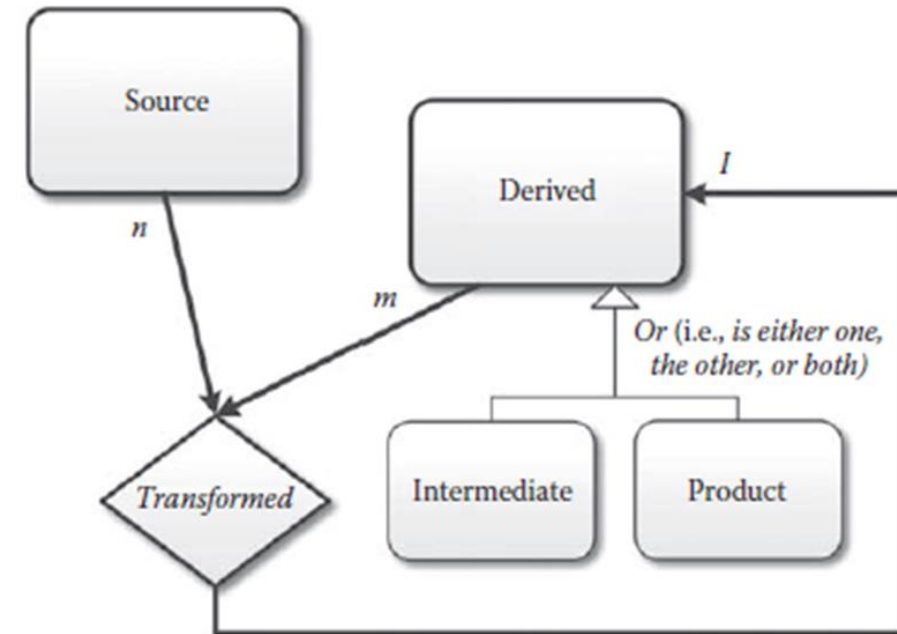
Data lineage vocabulary helps communicate how data is processed in an information system

and can aid thinking about how to meet privacy by design requirements
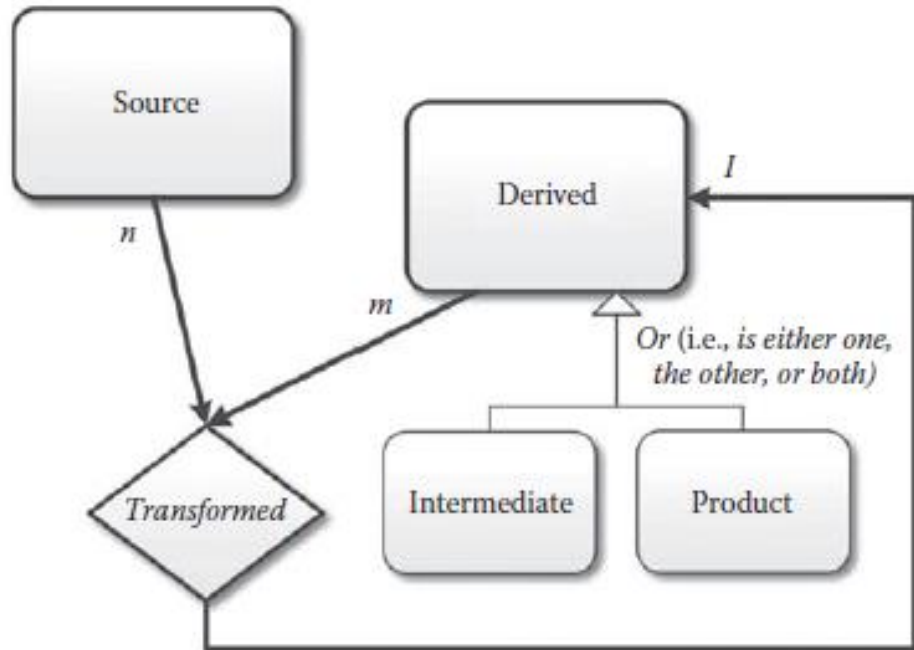
Source datasets *may contain personal data*

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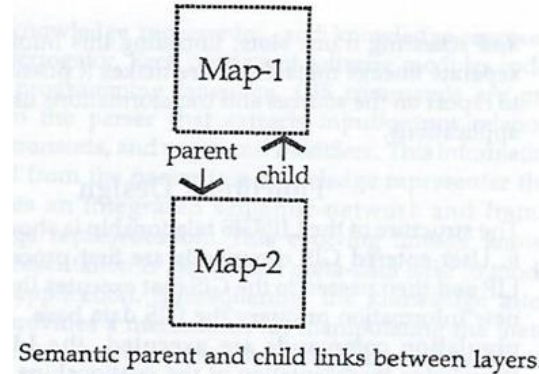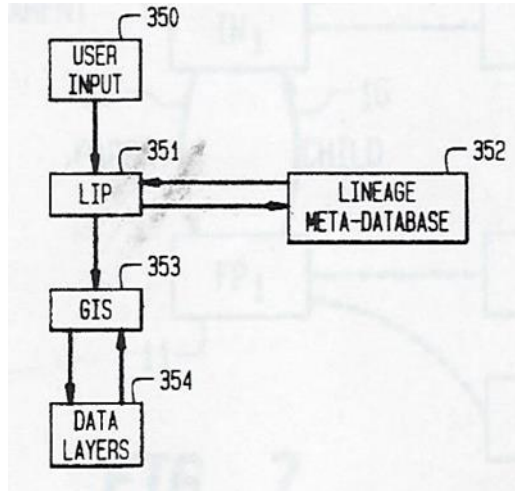
Input datasets provided with parent links pointing to output datasets can answer the question: ***Who am I the parent of?***

Output datasets' child links connect them back to their input datasets can answer the question: ***Who am I the child of?***
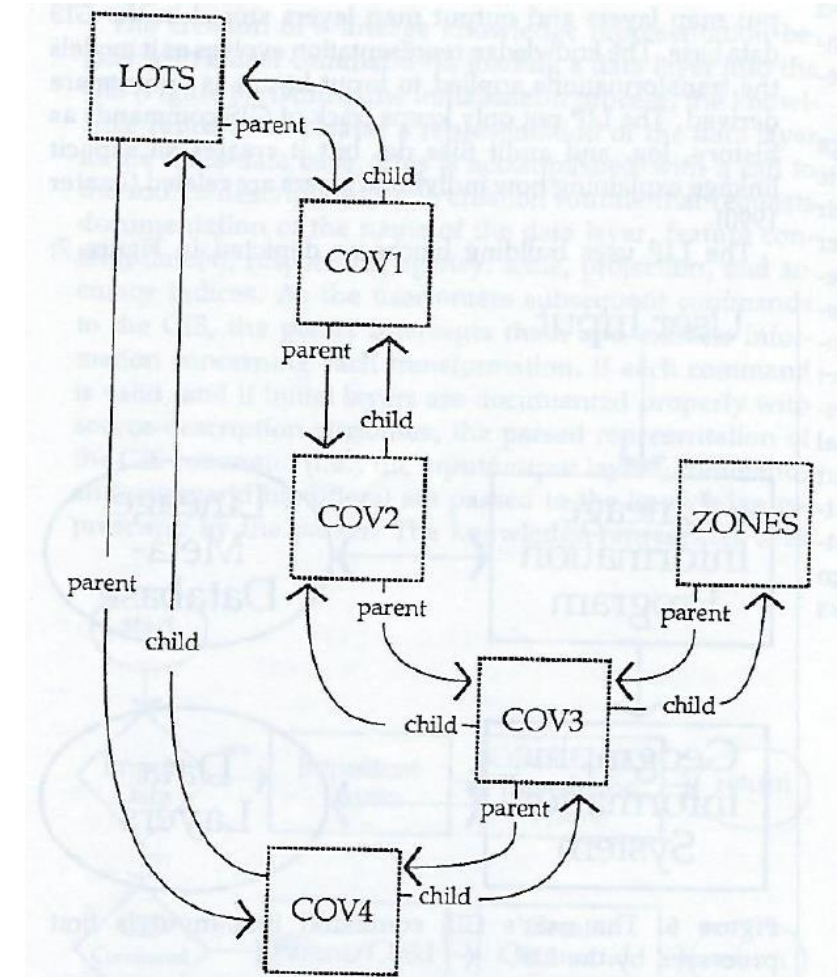
Descendants function traces parent links to identify all datasets derived from a source or other derived input dataset used within the application.

```
(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)
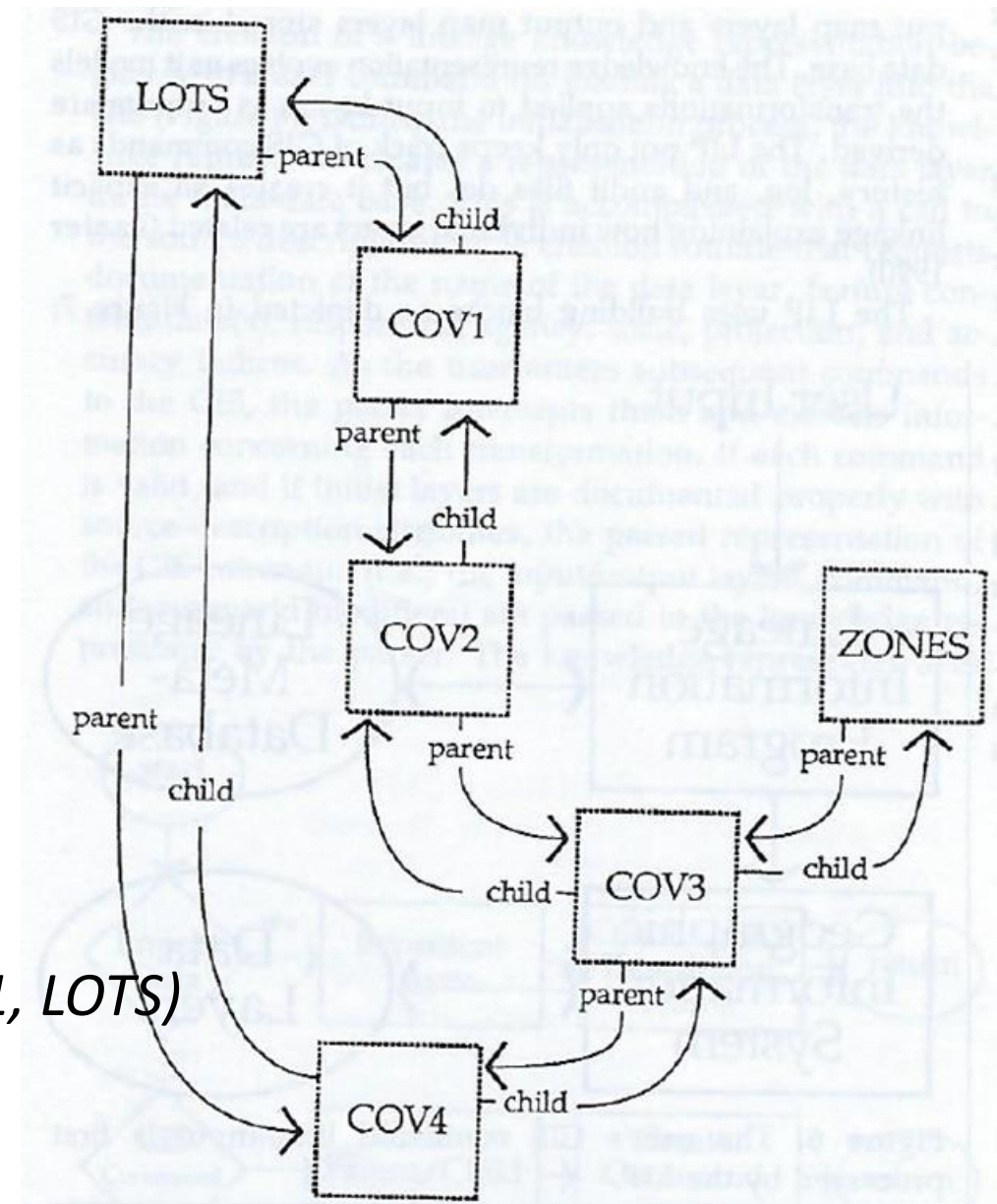
Source properties can include:

- 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,…
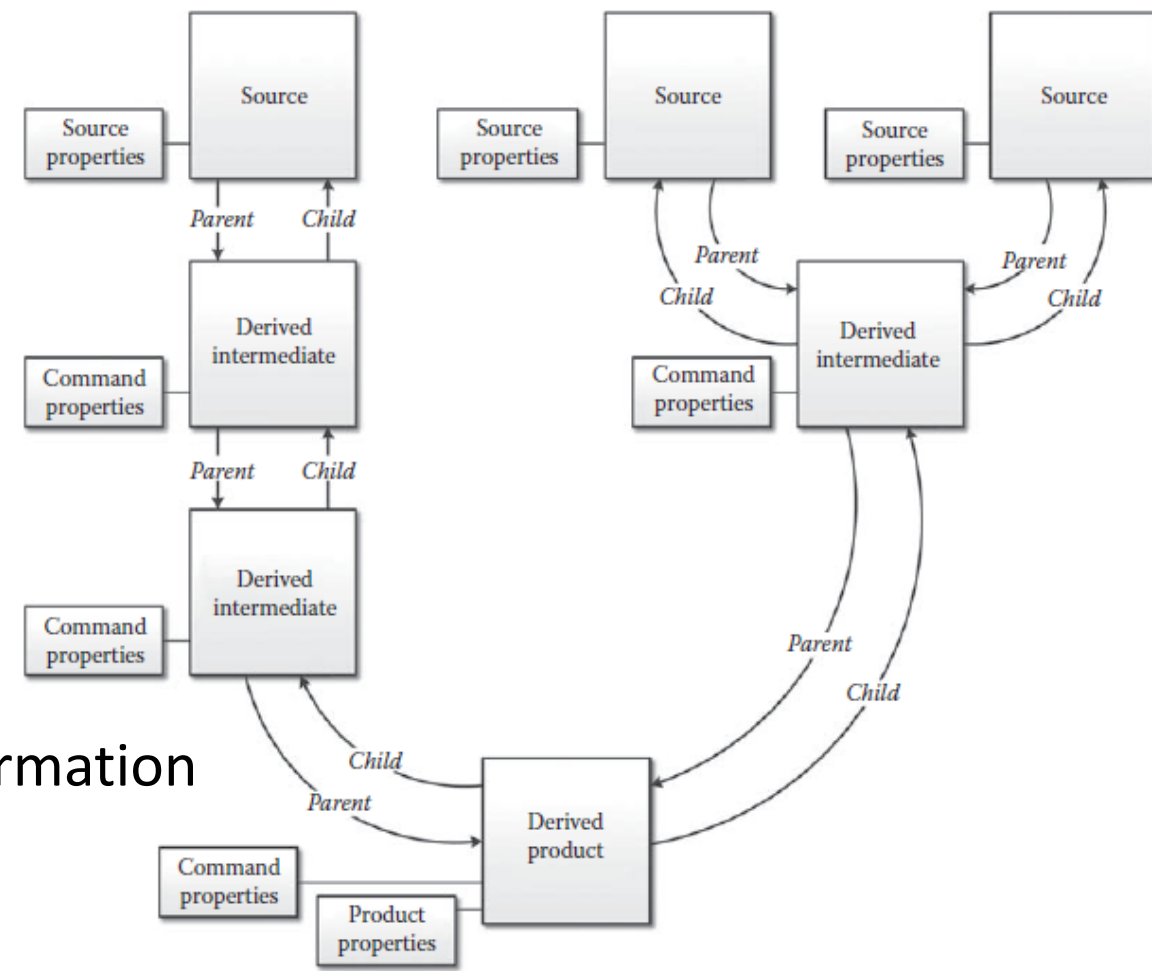
Meet Geo_lineus
source metadata input


```
(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!
```
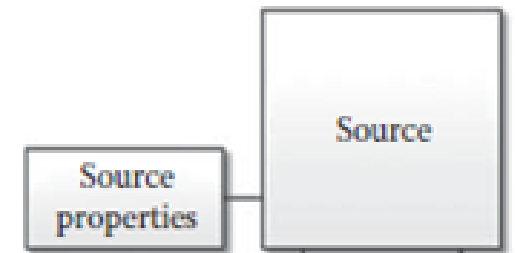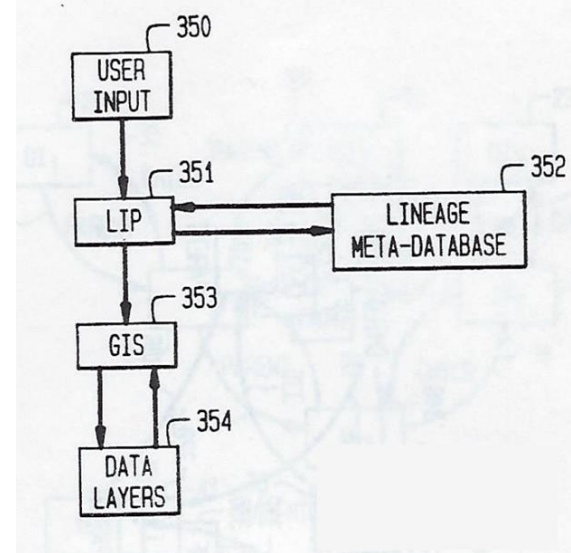

SOURCE DESCRIPTION FRAME	
SOURCE:	Digital line graph
FEATURES:	Hydrography
S_DATE:	4/7/83
AGENCY:	USGS
SCALE:	1:100,000
PROJECTION:	Mercator
ACCURACY:	+-10 meters Horiz

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)
```
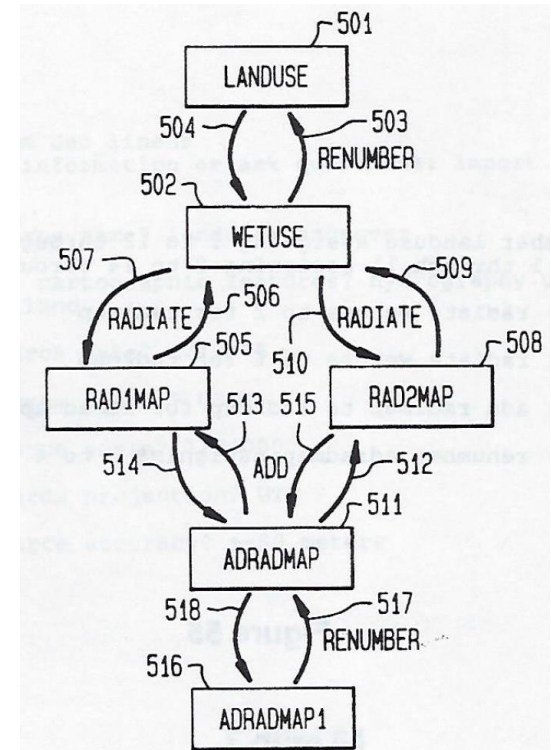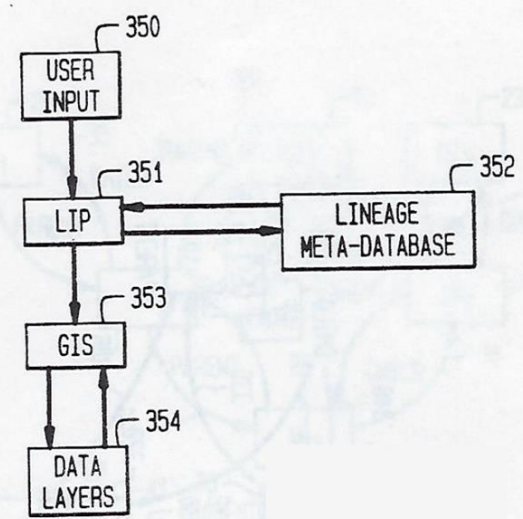
Product Metadata input…
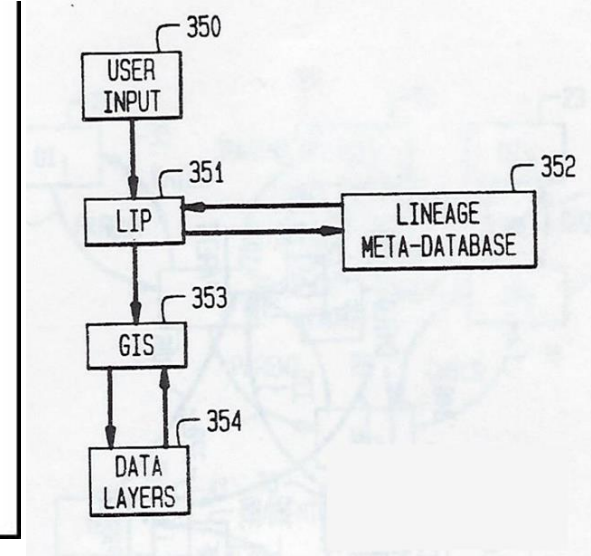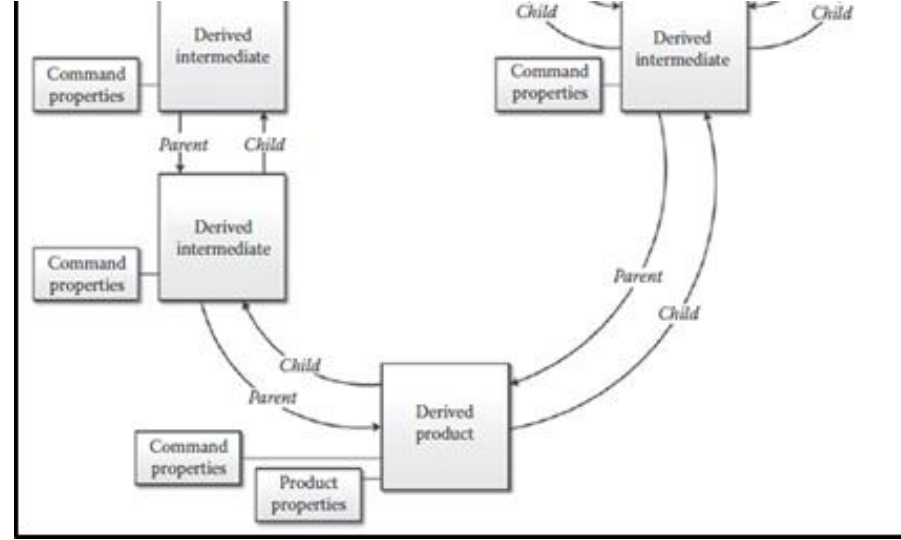

```
     export cover adradmap1 eco_zones

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!
```
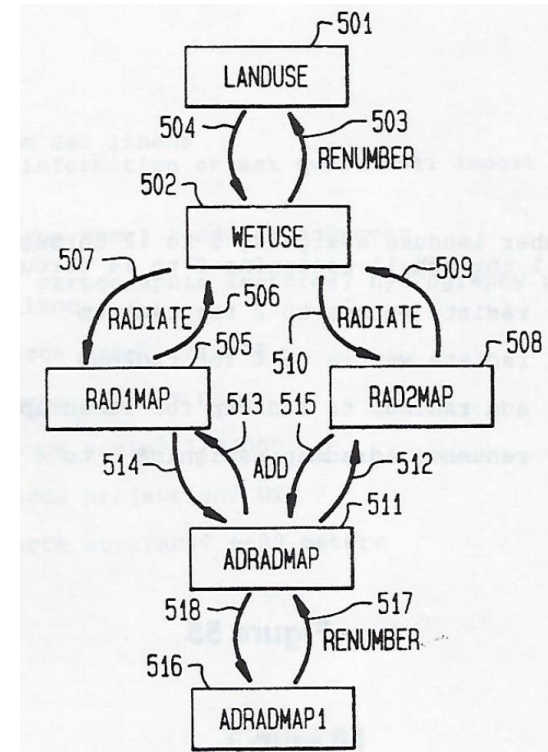
Querying metadata...
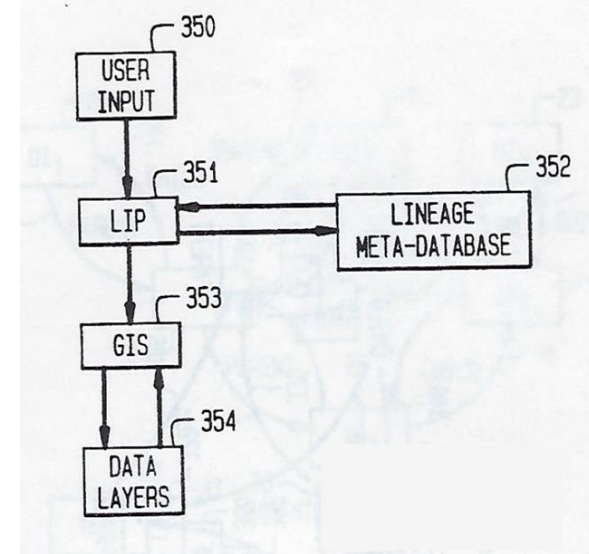
Is landuse a parent of adradmap

(YES INDEED LANDUSE IS A PARENT OF ADRADMAP)
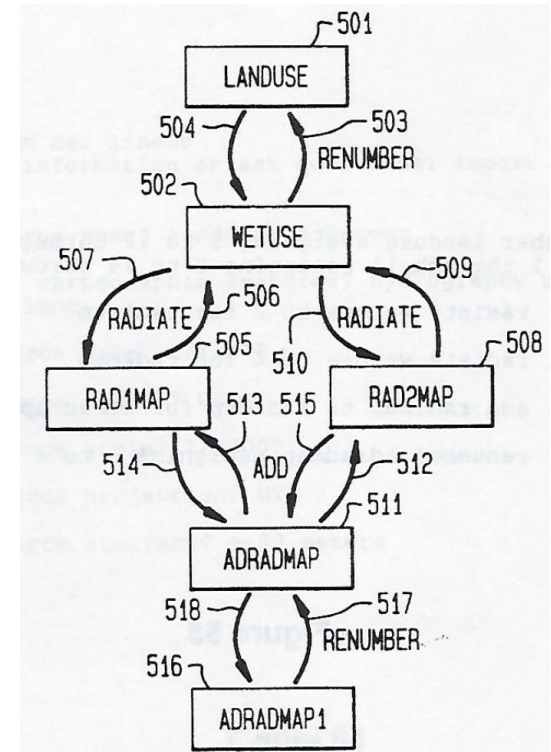
Querying metadata...
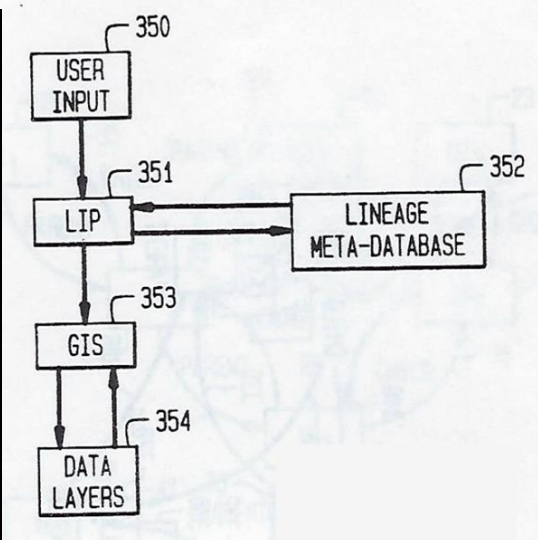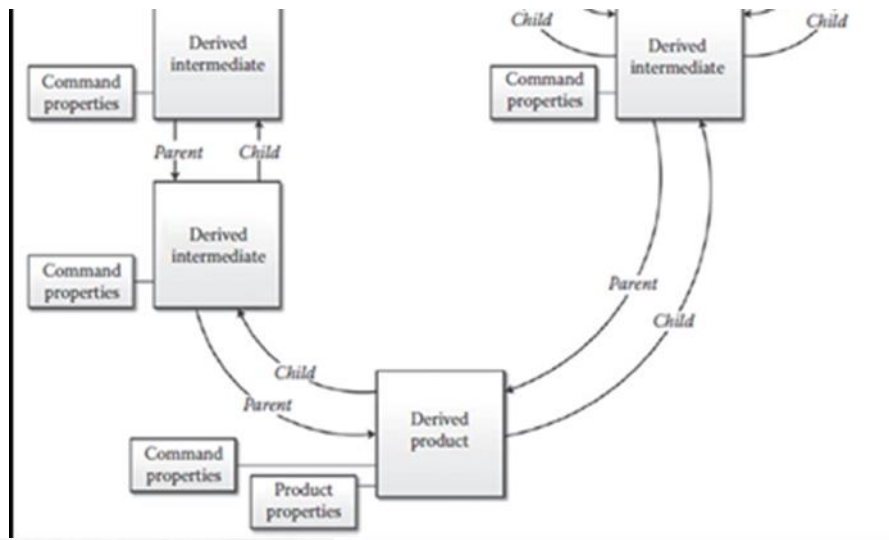
What is the lineage of adradmap1

(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)
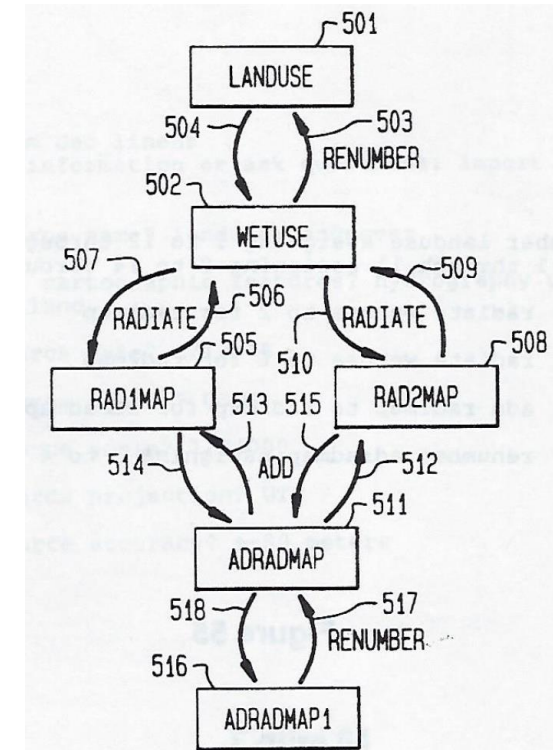
Querying metadata...

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)
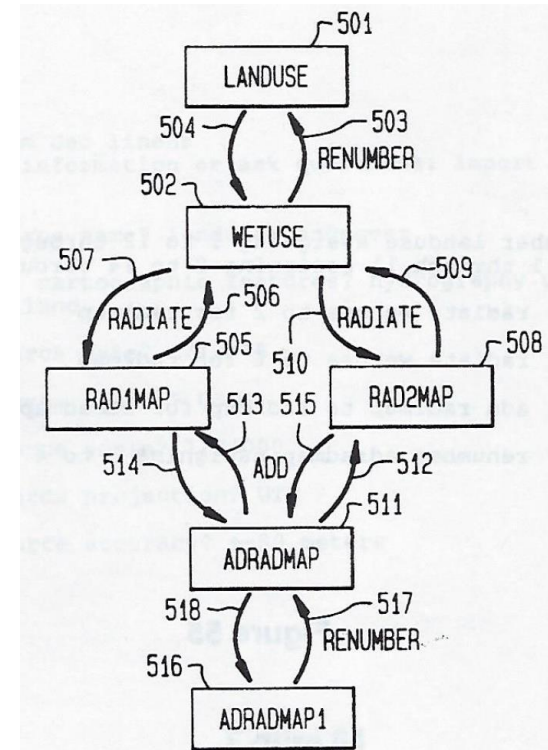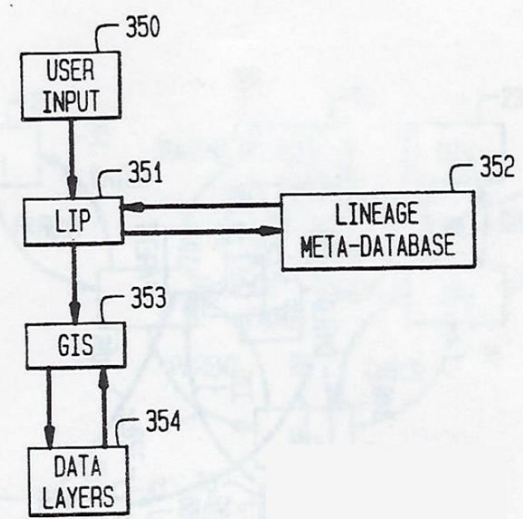
Adding a graphical user interface…

GUI design by Rupert Essinger

Working with source and command metadata

Update propagation…

Data source metadata based integrity constraint

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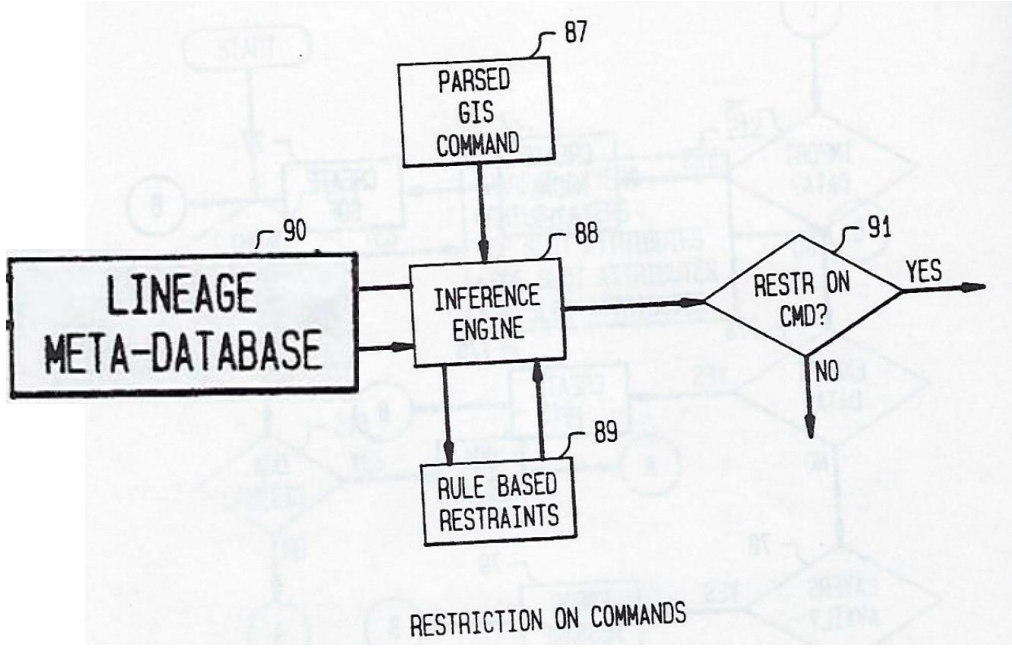
```
(setq intersect_rules
    '((rule intersect1
        (if (not (equal (scale inmap)
                    (scale intersectmap))))
            (then ("INPUT SCALES NOT EQUAL" )))

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

Lineage Window

INPUT PROJECTIONS NOT EQUAL
Reproject one of the maps.

COV2

GIS Window

```
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]
```
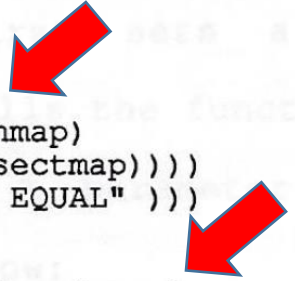
Conclusion:
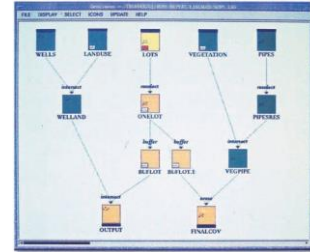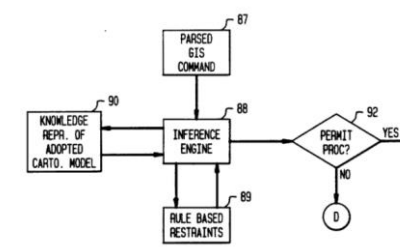
Data lineage metadata can help information system developers meet key data protection by design requirements:

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**

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

Agenda

✓ Online privacy

✓ Privacy and data protection by design

✓ *...with data provenance and lineage metadata*