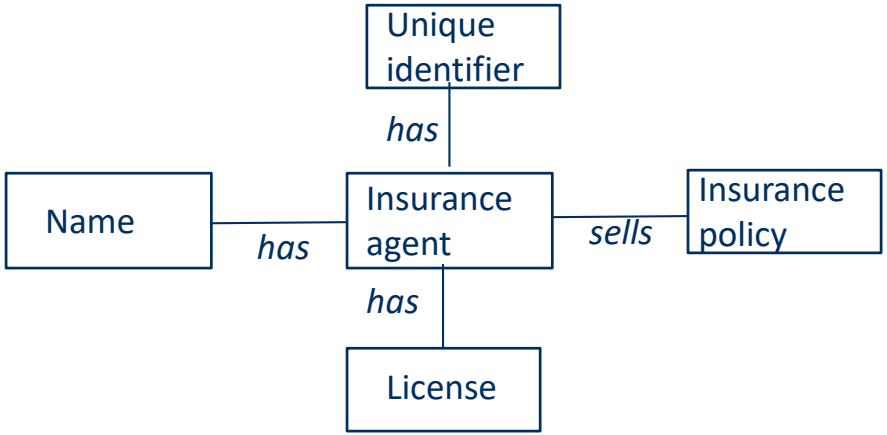
OBIDAM: Data Stewardship, Part 2

Erik Perjons

DSV, Stockholm University

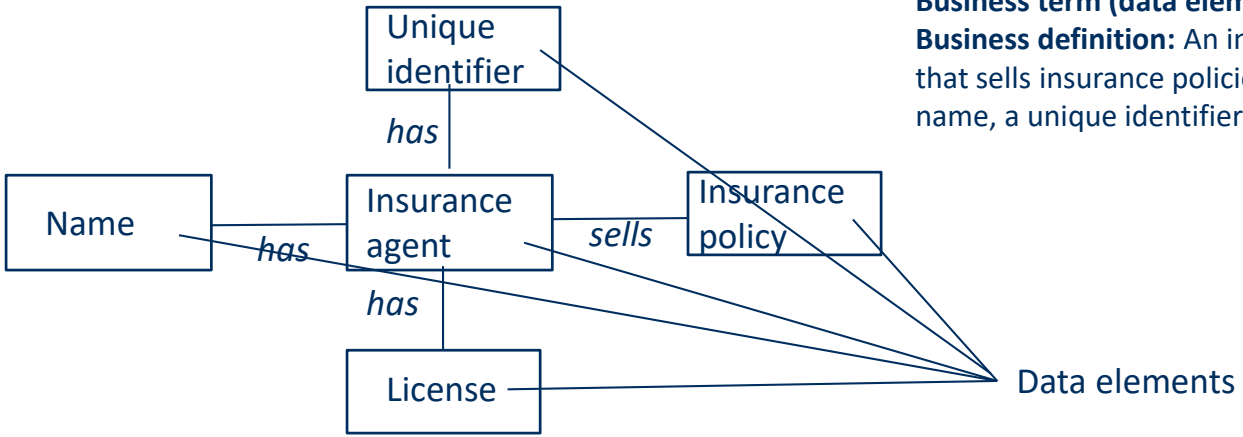
Data element – what does it mean?



Business term (data element): Insurance agent
Business definition: An insurance agent is an individual that sells insurance policies. The agent needs to have a name, a unique identifier and a licence

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Data element – what does it mean?



Business term (data element): Insurance agent
Business definition: An insurance agent is an individual that sells insurance policies. The agent needs to have a name, a unique identifier and a licence

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Business stakeholder

- A business stakeholder (in the book of Plotkin) is a role that will be impacted by a decision regarding data and data-related issues
- A business stakeholder in data governance is often a user of the data

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

What does a data steward needs to do?

Stockholms universitet

- Determin key business data elements
- Assign key business data elements to data stewards
- Define key business data elements
- Specify business rules for creation and usage of key data elements
- Specify derivations rules for key data elements
- Measuring data quality of key data elements
- Specify data quality rules of key data elements
- Profiling data quality

What does a data steward needs to do?

Stockholms universitet

- Specify processes for data governance/stewardship
- Establish procedures for carrying out steps in the processes in practice
- Create and manage quality metadata (e.g. business definitions and business rules for creation and usage, etc)
- Introduce and manage tools, such as business glossary, metadata repository, data profiling tool and issue log
- Develop a communication plan

Determin key business data elements

Stockholms universitet

- **Determin key business data elements** - The first step for a data steward is to decide which business data elements are key to bring under governance
- That is, **identifying business data elements that are worth spending time on to govern**, by doing an informal ROI analysis/cost benefit analysis

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Determin key business data elements

Stockholms universitet

- Key business data elements candidates:
  - **Data elements used in financial reporting** – that is, data that is reported to the financial community
  - **Data elements for complience and regulatory** – that is, data that need to be reported due to regulations
  - **Data elements introduced by the executives** – that is, data that are often found in executives' presentations
  - **Data elements used by high-profile projects**
  - **Data elements decided by the data steward** – that is, based on her/his experiences

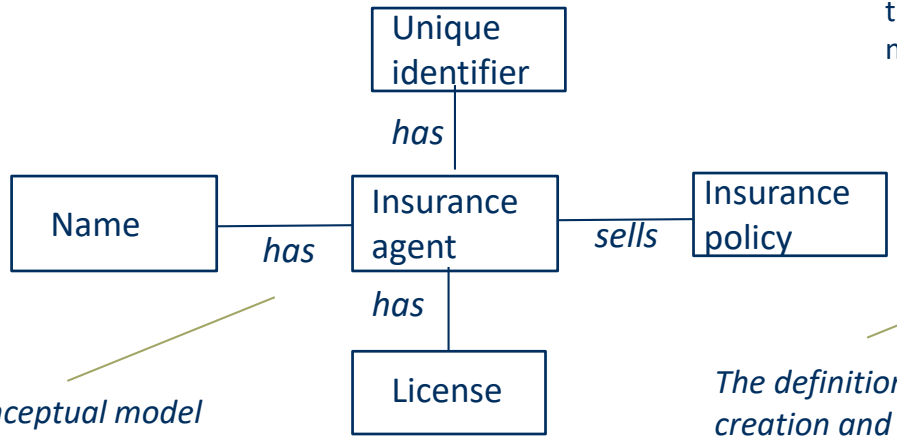
[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Assign responsibility to key business data elements

- Every key data element should also be owned by one business function, and be governed by one data governor (decided by the data governance board)
- If a key data element is used in two or more business functions, the following guidline can be applied:
  - Data elements should be owned by the business function:
    - that would fundamenally change if the definition of the data element change
- Every key data element should have one responsible data steward (decided by the Data stewardship council)

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Define key business data element and specify creation and usage business rules



Business term (data element): Insurance agent
Business definition: An insurance agent is an individual that sells insurance policies. The agent needs to have a name, a unique identifier and a licence

Creation business rules:
- Must assign a unique identifier
- Must have a valid licence
- Must have a valid address

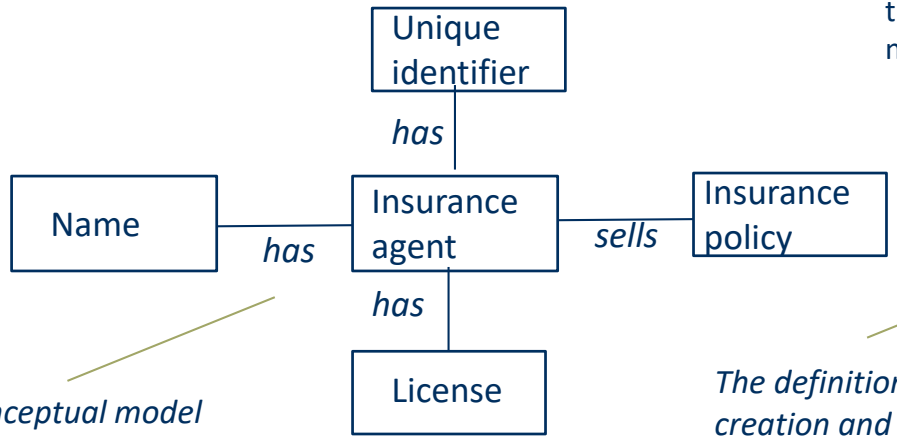
Usage business rules:
- Can only be used for insurance policies that the insurance agents are assigned to
- The insurance agent' licence must be valid when used (validy test must be carried out)

A conceptual model can be used to visualize the definition of the term by relating it to other defined terms

The definitions and creation and usage business rules of a business term are part of the business glossary

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Define key business data element and specify creation and usage business rules



Business term (data element): Insurance agent
Business definition: An insurance agent is an individual that sells insurance policies. The agent needs to have a name, a unique identifier and a licence

Creation business rules:

- Must assign a unique identifier
- Must have a valid licence
- Must have a valid name

Usage business rules:

- The insurance agent's licence must be valid when using an instance of "insurande agent" for closing a deal with a customer

A conceptual model can be used to visualize the definition of the term by relating it to other defined terms

The definitions and creation and usage business rules of a business term are part of the business glossary

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Data definitions - requirements

- The definition should be expressed in a business language
- The definition should be intensional – that is, express the meaning or content of the term
- The definition should be specific enough so that users should be able tell how the term differs from similar terms
  - Example: The term ”financial transaction” was broken up in two terms ”attemped financial transaction” and ”completed financial transaction”
- The words used in the definition should, if possible, be linked to already defined terms (see conceptual model that visualizes this)

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Specify business rules for creation of data elements

- The business rules for creation specify under which conditions an instance of a data elements can be created
- These conditions could include:
  - At which point in the business process the data instance must be created
  - Which other data must be available prior to creating the data instance (can also be shown using multiplicity in the conceptual model)
  - What business function is allowed to create the data

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Specify business rules for usage of data elements

- The usage business rules specify how an instance of a data element is allowed to be used
- These rules could include:
  - Relationships to other data that must exist
  - Which business process must use the data and in which way each process use the data
  - Validy test that must be applied

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Specify derivations rules for data elements

- Values of data elements instances can be derived. For example, the value of data element ”age” is derived (calculated) from the value of the data element ”date of birth” and the value of the data element ”today’s date”
- Another example of a derivation is: A person’s status move from the value ”prospect” to ”customer” if the person open up a ”bank account”
- Such derivations – called ”derivation rule” - needs to be defined

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Specify processes for data governance/stewardship

- Processes specified for data governance and data stewardship make everybody aware of how job should be carried out – these processes are owned by the data governance council but data stewards will have a say in the design of them
- For example, these processes describe how the data steward shall interact with other roles and different councils that are part of the data governance program
- For describing these processes, business process models can be used
- Business process tools (e.g, a workflow or case management system) can also be used for shepard the processes through the required steps

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Measuring data quality

- Data quality needs to be measured against a set of requirements – that is, data quality rules - and these requirements (read: data quality rules) are based on how the organization wants to use the data
- That is, in order to understand the data quality of an organization it needs to:
  - specify data quality rules – specify what data quality means
  - introduce a data profiling process, using a data profiling tool, for measuring data quality of existing data in the organization
  - measure the quality of the data in the organization using the data quality rules and a the data profiling process and tool

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Measuring data quality

- If data quality is not measured, the organization may not have a full understanding of the quality of the data, which may lead to incorrect decisions, wrong investments, accidents, violation of regulations/compliance, loosing credibility among customer and clients

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Specify data quality rules

- Data quality rules should consist of two parts:
  1) a **business statement of the rule** - explains what quality means in business terms, and
  2) a **data quality rule specification** - explains what is good quality at the physical data store level
- An example of a data quality rule:

  Business statement: The marital status code for employees shall have values of single or married. It shall not be left blank. A value must be selected in the system when entering a new employee in the system.
  Data quality rule specification: EmployeeDemographics.Martial_Cd shall be ”Sng” or ”Mard”. Blank is considered as an invalid value.

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Specify data quality rules

- The **data quality rule specification** can be categorized in three main types:
  - Single column content rules
  - Cross-column validation rules (typically in a single table)
  - Cross-table validation rules

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Data quality dimensions 1(3)

- What do we mean by data quality?
- What do we actually measure?
- The answers to these questions you need to specify a set of so called data quality dimensions, such as “accuracy”, “completeness”, and “timeliness”, each require different ways to measure data quality

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Data quality dimensions 2(3)

Examples of data quality dimensions:

- “Accuracy" means the degree to which data correspond to known correct values in the real world, provided by a recognized source of truth
- “Completeness" means the degree to which data are populated in a system – given the requirement stated in quality rules
- “Timeliness” means the degree to which data are available within the timeframe required by the business

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Data quality dimensions 3(3)

More examples of data quality dimensions:

- “Coverage" means the degree to which data provide business function with the data needed to carry out necessary functions
- “Validity" means the degree to which data conforms with acceptable content, stated in quality rules, such as format, valid value list, range, domain, data type etc
- “Uniqueness” means the degree to which data is allowed to have duplicated values
- “Integrity" means the degree to which data contain consistent content across multiple data sources

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

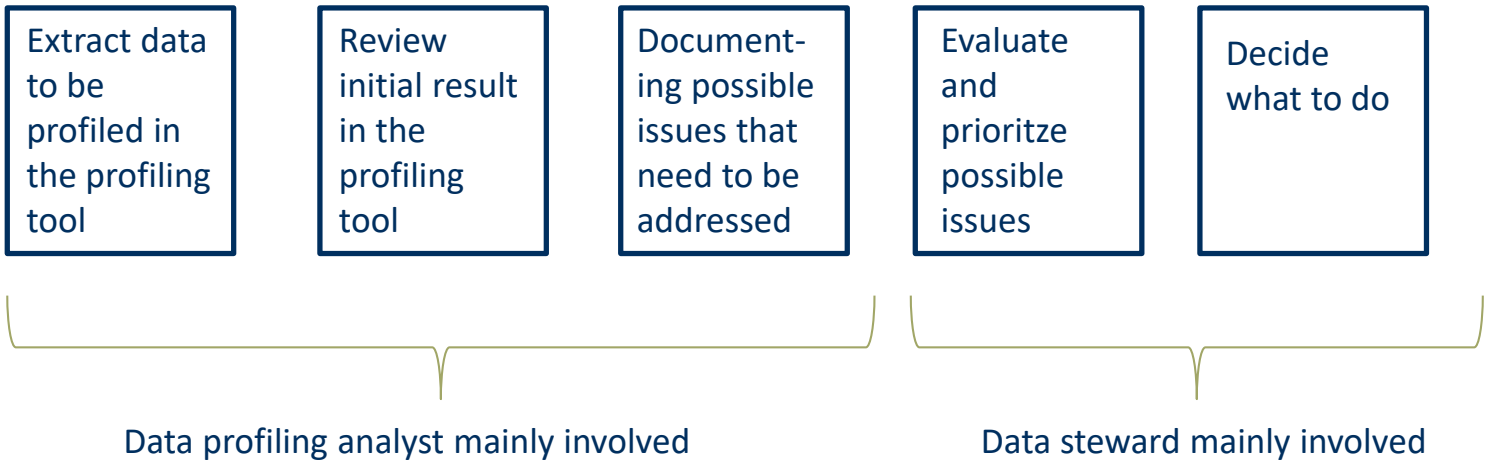
What is the right level of data quality?

- When has the level of quality been improved enough?
- In order to answer that question you need to decide the level of data quality, by, for example, introduce some kind of business cases (such as ROI or cost-benefit analysis)
- That is, the level of quality for the intended purpose need to be decided

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Data profiling process

- A data profiling process needs to be introduced for measuring the data quality. The process requires a data profiling tool.



[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Data profiling: Decide what to do

- The data steward can decide to the following in the last step in the process:
  - Decide if there is a need to evaluate the low data quality's impacts on the business
  - Decide if it is worth correcting low data quality
  - Decide if there are some workarounds that can be used and should be used (for example, using a ETL process to correct the data)
  - Decide if you need to identify the cause to the low quality
  - Decide if you need to set up automated enforcement of the data quality rules as well as notification

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Tool: Data profiling tool

- A data profiling tool can be used to measure the conformance to data quality rules – also called “goodness to fit”
- A data profiling tool can also be used to discover possible data quality rules, for example, the tool detect that a column in the database always has a unique value, therefore, the tool suggests that that column should have a uniqueness data quality rule

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Enforcing data quality 1(2)

- One important action for data stewards is to create incentives for data producers to be accurate and complete, for example, do not left blanks, even if the accurateness and completeness do not directly benefit the data producers themselves
- Another important action: Inform data producers and user not to correct data qualiy themselves – instead inform the data steward

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Enforcing data quality 2(2)

- During data load – from one data source to another - there different way to manage data that violate data quality rules, that is, one of the following approaches can be selected:
  - Separate data violating and data not violating the data quality rules
  - Fix the data that are violating the data quality rules
  - Document the data violating the data quality rules and inform the decision maker or keep a log of data that are violating the data quality rules

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Major reasons for low data quality

- Data producers are not accurate and complete when producing data, for example, they may leave blanks and do not change default values when needed
- Data quality rules are not enforced during data loads
- Data users are making their own corrections when identifying data quality issues, instead of informing a responsible for data quality (for example, a data steward that will enhance the data quality also for downstream users).

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Major reasons for low data quality

- Data quality rules are not defined
- Data quality is not measured (using data profiling process and a tool)
- Root causes to data quality are not fixed

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Tool: Issue log

- An issue log is an log where problems/challenges/questions about governed data are documented as well as information about the requester, the responsible, what actions has been taken, the actual status of the problems/challenges/questions, the priority among the actions, and, if there exist, the resolutions

.

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Tool: A business glossary

- A business glossary is where the business metadata is published, such as:
  - data definitions of the data element
  - derivation rules of the data element
  - creation and usage business rules
  - data quality rules/quality requirements
  - data security classifications
  - info of responsible data stewards
  - info of owning business functions (and data governors)
  - applications using the data element

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Tool: A business glossary

- A business glossary should have an effective search function
- A busines glossary should warn for duplicates

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Tool: A metadata repository

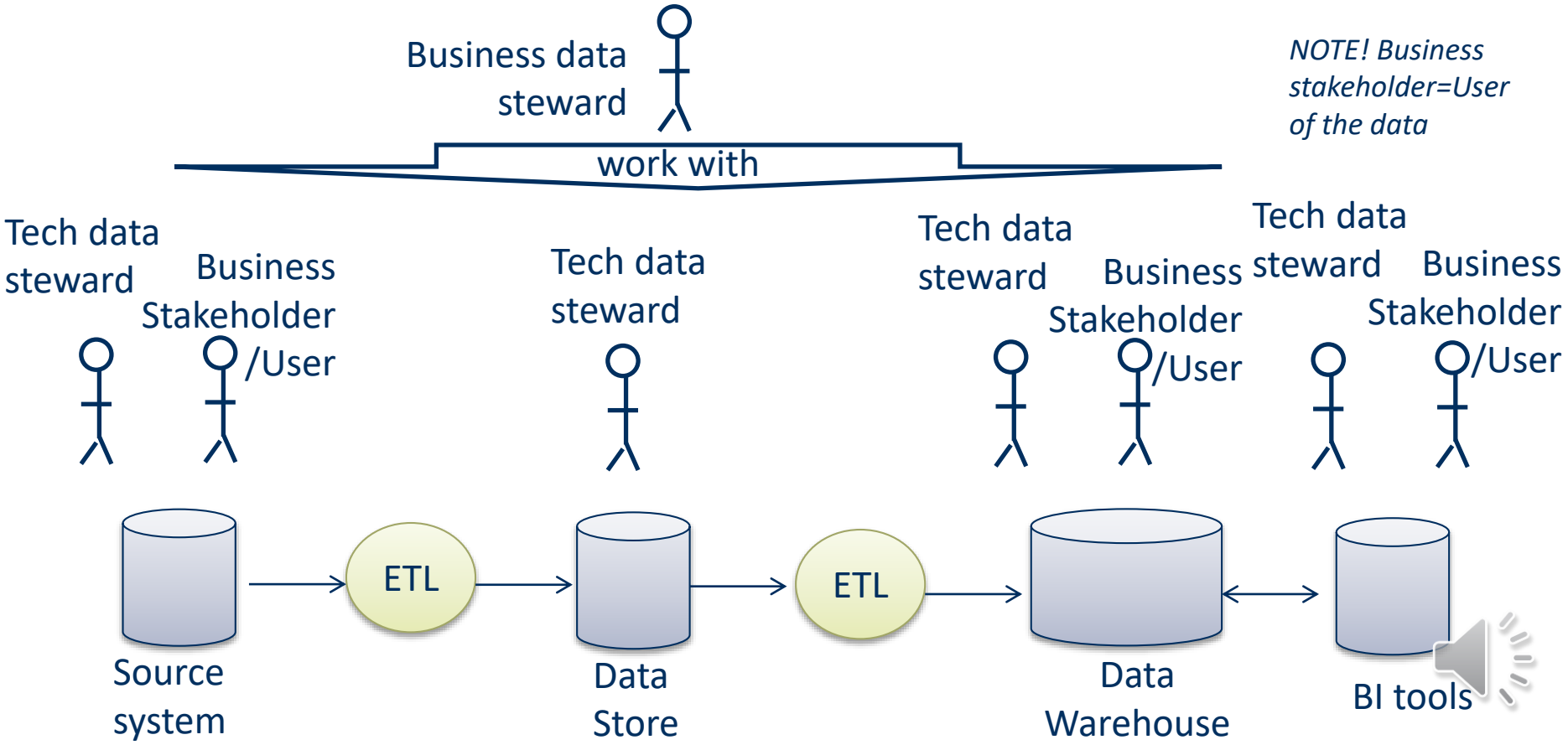
- A metadata repository is usually focused on physical and technical metadata, such as data models, database structures, metadata related to BI tools, ETL
- A metadata repository provide a link between the business data elements and the potentially many physical implementations of these elements
- A metadata repository can also describe the information chain – where data is created and how it is changed/manipulated and used downstream
- Business metadata is usually stored in the business glossary, but can also be part of the metadata repository

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Manage data in a information chain

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

NOTE! Business stakeholder=User of the data



Information chain

- **An information chain** - is the flow of data from the source to various targets where the data is used
- **An information chain** - is a supply chain for data, that is where data is created, modified and stored and where it is used in the chain
- Data must be manage across the information chain. **Upsteams** means earlier in the information chain and **downstreams** is later in the in the information chain

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

Tool: Communication plan

- The communication plan describes how major decisions need to be communicated in the organization. The communication plan needs to have information about:
  - what type of decision exists
  - how should the different types of decisions be communicated
  - who are the audience for the different types of decision
  - which medium shall be used for the different types of decisions
  - who is the responsible for communicating the different types of decisions

[Plotkin D. (2014) Data Stewardship, Morgan Kaufmann Publishers]

More to do for data steward

- Managing reference data
- Managing master data, including identity resolution
- … to be discussed in Week 3

.