cluedin

white paper

# How does the Data Steward and Data Citizen play a role today?

2019 | cluedin.com

# How does the Data Steward and Data Citizen play a role today?

There are new roles emerging in the data space, but yet very few businesses have established a process to enable these new roles. The new roles revolve around this idea of being responsible for data quality of a particular set of data. The CRM is probably the best example in the past, of systems that are governed the most - as, if the CRM is of good quality, then proper sales forecasting can be achieved and more.

***To know if you have these types of people at your work already, ask yourself this question, if the data quality of a system is bad, do you know who is responsible? Could this person be held accountable or even fired if the quality is not good enough or if initiative is not shown to make it better over time?***

Even with this organisational governance in place, few have the technology needed to fuel these individuals with a good process around increasing the quality of data. A lot of this comes from the fact that this person may not know the context behind the data, they may not be the best person to qualify data. Hiring a data expert for your team will only fix certain things, but what about data that needs context e.g. How would a new team member know if “Eric Smith” is actually spelt “Erik Smith”? However this would be glaringly obvious to a sales person that speaks to Erik on a daily basis.

***The problem with the data steward role today is that there is not a good process around what data they really should be “fixing” and what is the responsibility of someone closer to the data itself.***

This is the main difference between the Data Steward and the Data Citizen. The Data Citizen is a business user, potentially slightly data savvy that is also close to the data. They will typically have access to the source system and are constantly within that system on a daily basis. They might not be the actual sales person themselves, but potentially someone who qualifies data and maintains a “clean” CRM. They are part of sales meetings, they know what deals are being worked on and what discussions are happening, but they might not be the person who actually is making the sale.

## It is important that the Data Steward is given an interface that is as simple as possible.

Unfortunately, the user interfaces to be able to address this problem do come with some natural complexity i.e. if they need to fix data, they are best off fixing it in the source (some would think). This is not the only role of a Data Citizen today. The Data Citizen is responsible for fixing the “context” of data. Why? Because no company that I can name started off with the highest fidelity of data storage i.e. the Graph.

For us to transfer our data to a Graph world from the typical relational world, we need to infer and add context to the data as to form a high integrity network of our data. This can only be added with a certain level of confidence and there is a large need for inference of relations of data instead of hard and firm rules. Below is a list of the types of stewards and a small description of their perceived responsibilities.

### Data object data steward

Responsible for managing reference data and attributes of one business data entity.

### Business data steward

Responsible for managing critical data, both reference and transactional, created or used by one business function.

### Process data steward

Responsible for managing data across one business process.

### System data steward

Responsible for managing data for at least one IT system.

This is why one of the upcoming roles of the Data Citizen is to label data and provide context, but not correct the data. Data correction can easily happen in the source system, context can not easily be added in all source systems, but systems like CluedIn specialise in attaching context to data.

Here is a good example, if you wanted to define relationships between data in a system like Salesforce, things would get out of hand very quickly. Why? Because Salesforce is backed by Relational Databases. Imagine that for a Contact, you wanted a field to be able to set all the relationships of this data to other data. What is the core problem? We are only looking at direct relationships. To solve many data challenges today, this doesn't scale and it doesn't even address the core problem. Let's just throw some very practical examples at you that could be the types of things you need to clean on a daily basis:

*Are Lego and Lego ApS the same company?*

*Does Erik still work for Lego?*

*I need all sales related emails sent to Lego.*

*I have a Gender field that shows Male, Mail, M, F, Female, Mand, Kvinde. Which value should I normalise this to?*

*I need all sales related emails sent to Lego.*

*Is "Senior Innovation Guide" a Job Title or is it a Document Name?*

*There is mention of an Erik Smith in a Word Document. Which Erik Smith is it? We know more than one.*

Under what context was this data entered? i.e. there is no reason you should treat old data like you treat it today e.g. If you were upset that a certain Sales Deal didn't have a close date, is that incomplete data or is it actually that at the time there was no policy to set this? Suddenly with this context in place, this data is maybe incomplete BUT it is valid. These are two different things.

- Who should fix the first question?
- What determines sales related emails?
- Who would know the answer to the 3rd question? What if they are a subsidiary?
- Why could this not be a Job Title?
- Whose job is it to normalise data into the company standards?

## So why is this, now, the right approach? Because technology has caught up.

Machine Learning is more available and ubiquitous today. What changes is the way we work, and this is the hard part. Imagine that you were in charge of Master Data Management for your large company. On a daily basis you would be setting up rules, policies, fixing data and flagging invalid data. Suddenly, instead of correcting data you are just telling a system that it did something wrong. Imagine how frustrating it would be to know how to fix the data, but there was not an option (I am of course just trying to make a point, of course you could manually fix it) to correct the data, but rather you simply had to tell a system that it did something wrong. In the future, you might even tell it a category of what it did wrong, but for now it is a simple binary option. Why on earth would you do this? It is all about scale. Imagine telling a system that "Heroku" is not the name of a person, then imagine it asking 5 more times and each time your answer is "no". In the normal world, if it later would want to determine if "Jetstar" was a person, instead of asking you, it has been trained to identify things like this on your behalf.

Before we go off and lose our minds, we can't forget that systems like CluedIn still allow you to set rules, this is simply for catering for the data that falls outside the rules.

- ☑ Has clear and unambiguous definitions of data elements.
- ☑ Does not conflict with other data elements in the metadata registry (removes duplicates, overlap etc.)
- ☑ Has clear enumerated value definitions.
- ☑ Is still being used (remove unused data elements)
- ☑ Is being used consistently in various computer systems
- ☑ Is being used, fit for purpose = Data Fitness
- ☑ Has adequate documentation on appropriate usage and notes
- ☑ Documents the origin and sources of authority on each metadata element
- ☑ Is protected against unauthorised access or change

## A question we often get asked, is how much training does a system need. Wait for the really annoying answer.

As much as you can give it. Annoying, but couldn't be more true. Imagine you are bringing up a child. Wouldn't you argue that the more you train it, the better it will come out? It is a generalisation, but I think we can meet somewhere in the middle. Imagine that you trained your child that a knife was bad and the teddy bear was good. If you do this for long enough, in essence, and you were to then show them a chainsaw (note: Don't give children a chainsaw), then without seeing it before, they should be able to decide that it is a bad thing.

See how this is different? You are not monitoring your child 24/7 and setting rules for "If Knife or If Saw or If ChainSaw then this is bad". You would agree that this would never scale right? Right. So why then are we doing this with data? Wouldn't you agree that the repercussions for something happening with a child is much higher than something happening with data? (Yes, in saying this, I know what some of you said Data).

Then why do we think this will scale with our data? It doesn't. Instead, it would be important for us to give our child (or data) guidance and then at some point we need to let it live. If you did a good job at training it, then on most occasions it will do something good, but there will be moments where it gets things wrong.

This is the beauty and the difference between a children and data. In data, we can actually watch it 24/7. There are no “privacy” issues with spying on our own data all the time. Hence we can always use rules to catch the data that strictly cannot have a fuzzy aspect to it and then put in training sets of data to train for that type of problem.

## Shall we complicate the situation now? Why not. So then what is the role of the Data Custodian?

- Access to the data is authorized and controlled
- Data stewards are identified for each data set
- Technical processes sustain data integrity
- Processes exist for data quality issue resolution in partnership with Data Stewards
- Technical controls safeguard data
- Data added to data sets are consistent with the common data model
- Versions of Master Data are maintained along with the history of changes
- Change management practices are applied in maintenance of the database
- Data content and changes can be audited

Turns out that the Data Citizen mentioned above is very similar to the Data Custodian, so we can immediately breathe a sigh of relief. Why? Because without a doubt, there will be a shortage of the roles that are necessary to fulfil the necessary data stewardship and everything that comes with it. We can definitely say that our customers are already well ahead of the curve. A lot of this comes from exposing them to this new world. Like the Data Science era, we will see a shortage of Data Stewards and it will essentially be a race.

cluedin